BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12Q 1/68, C12P 19/34

A1

(11) International Publication Number: WO 93/17126

(43) International Publication Date: 2 September 1993 (02.09.93)

(21) International Application Number: PCT/US93/01552

(22) International Filing Date: 19 February 1993 (19.02.93)

(30) Priority data:
07/838,607 19 February 1992 (19.02.92) US

(71) Applicant: THE PUBLIC HEALTH RESEARCH INSTITUTE OF THE CITY OF NEW YORK, INC. [US/US]; 455 First Avenue, New York, NY 10016 (US).

(72) Inventors: CHETVERIN, Alexander, B. ; 24 Block "AB", #238, Pushchino, Moscow, 142292 (RU). KRAMER, Fred, Russell ; 561 West 231 Street, Riverdale, NY 10463 (US).

(74) Agents: JACOBS, Seth, H. et al.; Davis Hoxie Faithfull & Hapgood, 45 Rockefeller Plaza, New York, NY 10111 (US).

(81) Designated States: AT, AU, BB, BG, BR, CA, CH, DE, DK, ES, FI, GB, HU, JP, KP, KR, LK, LU, MG, MN, MW, NL, NO, PL, RO, RU, SD, SE, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, SN, TD, TG).

Published
With international search report.

(54) Title: NOVEL OLIGONUCLEOTIDE ARRAYS AND THEIR USE FOR SORTING, ISOLATING, SEQUENCING, AND MANIPULATING NUCLEIC ACIDS

(57) Abstract

The present invention relates to new oligonucleotide arrays and methods of using oligonucleotide arrays. Binary oligonucleotide arrays, having binary oligonucleotides characterized by a constant nucleotide sequence adjacent to a variable nucleotide sequence, are used for sorting and surveying nucleic acid strands. Oligonucleotide arrays are used for sorting mixtures of nucleic acid strands, making immobilized partial copies of nucleic strands, ligating strands, or introducing site directed mutations into strands. Information is obtained for determining the sequence of a nucleic acid strand, alone or in a mixture, by generating partials of the strand and, for groups of partials having the same terminal variable oligonucleotide, separately determining the presence and sequence of all variable oligonucleotides. Arrays are also used to order previously sequenced nucleic acid fragments and to allocate ordered allelic fragments to chromosomal linkage groups.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	MR	Mauritania
AU	Australia	GA	Gabon	MW	Malawi
BB	Barbados	GB	United Kingdom	NL	Netherlands
BE	Belgium	GN	Guinea	NO	Norway
BF	Burkina Faso	GR	Greece	NZ	New Zealand
BG	Bulgaria	HU	Hungary	PL	Poland
BJ	Benin	IE	Ireland	PT	Portugal
BR	Brazil	IT	Italy	RO	Romania
CA	Canada	JP	Japan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo			SE	Sweden
CH	Switzerland	KR	Republic of Korea	SK	Slovak Republic
CI	Côte d'Ivoire	KZ	Kazakhstan	SN	Senegal
CM	Cameroon	LI	Liechtenstein	SU	Soviet Union
CS	Czechoslovakia	LK	Sri Lanka	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	MC	Monaco	UA	Ukraine
DK	Denmark	MG	Madagascar	US	United States of America
ES	Spain	ML	Mali	VN	Viet Nam
FI	Finland	MN	Mongolia		

NOVEL OLIGONUCLEOTIDE ARRAYS AND THEIR USE FOR SORTING, ISOLATING, SEQUENCING, AND MANIPULATING NUCLEIC ACIDS

Field of the Invention

This invention is in the field of sorting, isolating, sequencing, and manipulating nucleic acids.

Background of the Invention

Ordered arrays of oligonucleotides ("oligos") immobilized on a solid support have been proposed for sequencing DNA fragments. It has been recognized that hybridization of a cloned single-stranded DNA fragment to all possible oligo probes of a given length can identify the corresponding, complementary oligo segments that are present somewhere in the fragment, and that this information can sometimes be used to determine the DNA sequence. Use of arrays can greatly facilitate the surveying of a DNA fragment's oligo segments.

In an oligonucleotide array each oligo probe is immobilized on a solid support at a different predetermined position. The array allows one to simultaneously survey all the oligo segments in a DNA fragment strand. Many copies of the strand are required, of course. Ideally, surveying is carried out under conditions to ensure that only perfectly matched hybrids will form. Oligo segments present in the strand can be identified by determining those positions in the array where hybridization occurs. The nucleotide sequence of the DNA sometimes can be ascertained by ordering the identified oligo segments in an overlapping fashion. For every identified oligo segment, there must be another oligo segment whose sequence overlaps it by all but one nucleotide. The entire sequence of the DNA strand can be represented by a series of overlapping oligos, each of equal length, and each located one nucleotide further along the sequence. As long as every overlap is unique, all of the identified oligos can be assembled into a contiguous sequence block.

There is an important limitation to sequencing by known surveying techniques. As relatively longer DNA strands are surveyed, there is an increasing probability that more than two

identified oligos will share the same overlapping sequence, i.e., the overlap is not unique. When this occurs, the sequence of the DNA cannot be unambiguously determined. Instead of one contiguous sequence block that contains the entire DNA sequence, the oligos can only be assembled into a number of smaller sequence blocks, whose order is not known.

Summary of the Invention

We have invented new oligonucleotide arrays and methods of using them.

A "binary array" according to the invention contains immobilized oligos comprised of two sequence segments of predetermined length, one variable and the other constant. The constant segment is the same in every oligo of the array. The variable segments can vary both in sequence and length. Binary arrays have advantages compared with ordinary arrays: (1) they can be used to sort strands according to their terminal sequences, so that each strand binds to a fixed location (an address) within the array; (2) longer oligos can be used on an array of a given size, thereby increasing the selectivity of hybridization; this allows strands to be sorted according to the identity of internal oligo segments adjacent to a particular constant sequence (such as a segment adjacent to a recognition site for a particular restriction endonuclease), and this allows strands to be surveyed for the presence of signature oligos that contain a constant segment in addition to a variable segment; (3) universal sequences, such as priming sites, can be introduced into the termini of sorted strands using the binary arrays, thereby enabling the strands' specific amplification without synthesizing primers specific for each strand, and without knowledge of each strand's terminal sequences; and (4) the specificity of hybridization during surveying can be increased by coupling hybridization to a ligation event that discriminates against terminal basepair mismatches.

A "sectioned array" as used herein is one divided into sections, so that every individual area is mechanically separated

from all other areas, such as, for example, a depression on the surface, or a "well". The areas have different oligos immobilized thereon. A sectioned array allows many reactions to be performed simultaneously, both on the surface of the solid support and in solution, without mixing the products of different reactions. The reactions occurring in different wells are highly specific due to the nucleotide sequence of the immobilized oligo. A large number of sortings and manipulations of nucleic acids can be carried out in parallel, by amplifying or modifying only those nucleic acids in each well that are perfectly hybridized to the immobilized oligos. Nucleic acids prepared on a sectioned array can be transferred to other arrays (replicated) by direct blotting of the wells' contents (printing), without mixing the contents of different wells of the same array. Furthermore, the presence of individual sections in arrays allows multiple rehybridizations of bound nucleic acids to be performed, resulting in a significant increase in hybridization specificity. It is particularly advantageous according to this invention to use a binary array that is sectioned.

Our invention includes methods of using sectioned arrays to sort mixtures of nucleic acid strands, either RNA or DNA. As used herein, "strand" means not just a single strand, but multiple copies thereof; and "mixture of strands" means a mixture of copies of different strands no matter how many copies of each are present. Similarly "fragment" refers to multiple copies thereof, and "mixture of fragments" means a mixture of copies of different fragments. The methods include sorting strands either according to their terminal oligo segments (3'-terminal or 5'-terminal), or according to their internal oligo segments on a binary array. Before or after sorting, universal priming region(s) can be added to the strands' termini to enable amplification. Binary sectioned arrays for sorting according to strands' terminal sequences ("terminal sequence sorting arrays") can be comprehensive. A "comprehensive array" is one wherein any possible strand will hybridize to at least one immobilized oligo. This type of sorting is particularly useful for preparing comprehensive libraries of fragments of a large genome. For example, in one

embodiment of the invention, strands of restriction fragments have their restriction sites restored and are sorted on a binary array. That array contains immobilized oligos whose constant segments contain the sequence complementary to the restriction site, and an adjacent variable segment. The array is complete, containing all variable sequences of each type in separate areas.

Our invention also includes using sectioned arrays for preparing every possible partial copy of a strand or a group of strands. The term "partial" refers to multiple copies thereof. Partials are prepared by either of the following methods: (1) terminal sorting on a binary sectioned array of a mixture of all possible partial strands generated by random degradation of a parental strand; or (2) generation of partials directly on an array, through the sorting on an ordinary sectioned array of parental strands according to the identity of their internal oligo sequences, followed by the synthesis of partial copies of each parental strand by enzymatic extension of the immobilized oligos utilizing the hybridized parental strands as templates. In either case, generated partials correspond to a parental strand whose 3' or 5' end is truncated to all possible extents (at the "variable" end of the partial), and whose other end is preserved (at the "fixed" end of the partial). These are "one-sided partials." Unless otherwise indicated the word "partial" is used herein to refer to one-sided partials.

Our invention also includes methods of using oligo arrays to obtain oligo information as part of a process for determining the nucleotide sequence of a long nucleic acid strand, or of many nucleic acid strands in an unknown mixture. A complete set of one-sided partials of the strand or strands is prepared on a sectioned array, and the oligo content of the partial strands in each well of the array is separately surveyed (i.e. each group of partials sharing the same oligo at the partials' variable end is surveyed).

Our invention also includes methods of using oligo arrays for ordering previously sequenced fragments from a first restriction digest of a large nucleic acid or even a genome.

Our invention also includes methods of using oligo arrays for allocating sequenced and ordered allelic fragments into their chromosomal linkage groups.

Our invention also includes a method of using binary arrays for surveying the oligos contained in strands or their partials. This method provides improved comprehensive surveys over the conventional surveying of oligos on an ordinary array.

Brief Description of the Drawings

Figure 1 shows a binary array.

Figure 1a shows an oligo immobilized in an area of a binary array.

Figure 2 shows a sectioned array having depressions.

Figure 2a shows a well of a sectioned array.

Figure 3 shows addition of a lattice to a support to make a sectioned array.

Figure 4 shows an example of sorting and amplification of restriction fragments on a sectioned binary array.

Figure 5 shows an example of preparing partials on a sectioned ordinary array.

Figure 6 shows, schematically, the order of steps for sequencing a complete genome.

Figure 7 shows, schematically, the use of a sheet with a number of miniature survey arrays for simultaneous surveying every well in a partialing array.

Figures 8 to 11 show examples of the determination of nucleotide sequences from indexed address sets obtained from analysis of mixtures of strands.

Detailed Description of the Invention

I. Oligonucleotide arrays

As used herein an "oligonucleotide array" is an array of regularly situated areas on a solid support wherein different oligos are immobilized, typically by covalent linkage. Each area contains a different oligo whose location is predetermined.

Arrays can be classified by the composition of their immobilized oligos. "Ordinary arrays" contain oligos comprised entirely of "variable segments". Every position of the oligo sequence in such a segment can be occupied by any one of the four commonly occurring nucleotides.

Comprehensive ordinary arrays are those wherein any segment of any possible strand will hybridize perfectly to the length of one or more immobilized oligos so that no strand is lost.

Binary arrays differ from ordinary arrays. A binary array is illustrated in Figures 1 and 1a. Figure 1 shows a substrate or support 1 having immobilized thereon an array of oligos 3, each oligo being in a separate area 2 of support 1. Figure 1a shows one area 2. A binary oligo 3 (many copies, of course) comprised of constant region 5 and variable region 6 is covalently bound to support 1 by covalent linking moiety 4.

Because of the constant segments, binary arrays provide means for the hybridization of longer sequences without increasing the size of the array. The constant segment can be located within the immobilized oligo either "upstream" of the variable segment (i.e., toward or at the 5' end of the oligo) or "downstream" from the variable segment (i.e., toward or at the 3' end of the oligo). The type of array that is chosen depends on the specific application. The constant region preferably is or includes a good priming region for amplification of hybridized strands by a polymerase chain reaction (PCR), or a promoter for copying the strand by transcription. Generally a length of 15 to 25 nucleotides is suitable for priming. The constant region can contain all or part of the complement of a restriction site. A binary array can be "plain" or "sectioned" (see below).

"Plain arrays" known in the art are arrays in which the individual areas are not physically separated from one another. Reactions carried out simultaneously are limited to those in which the nucleic acid templates and the reaction products are bound in some manner to the surface of the array to avoid the intermixing of products.

"Sectioned arrays" are divided into sections, so that each area is physically separated by mechanical or other means (e.g.,

a gel) from all the other areas, e.g., depressions on the surface, called a "well". There are many techniques apparent to one skilled in the art for preventing the exchange of materials between areas; any such method can be used to make a "sectioned" array, as that term is used herein, even though there might not be a physical wall between areas.

One type of sectioned array is illustrated in Figures 2 and 2a. Figure 2 shows a support sheet 60 having an array of depressions or wells 62, each containing many copies of an immobilized oligo 64. Figure 2a shows one well 62 of the array of Figure 2. Well 62 formed in support 60 has therein oligo 64 covalently bound to support 60 by covalent linking moiety 66. In practice one may prepare a plain array, e.g., on a flat sheet, and then, at a point during a series of steps involving its use, convert the array into a sectioned array, e.g., by making physical depressions in a deformable solid support to isolate the individual areas. The sectioned array can also be created by applying a lattice to the solid support and bonding it to the surface so that each area is surrounded by impermeable walls. An exploded perspective view of such a sectioned array is shown in Figure 3. Support or substrate 70, here a planar sheet, has mounted thereon and affixed thereto a lattice 72 comprised of a series of horizontal members 74, 76. The lattice members define a series of open areas which, in conjunction with support 70, define an array of wells 78. In some applications it is preferable to utilize a detachable lattice (or a removable cover sheet), so that the sectioned array can be converted back to a plain array.

Sectioned arrays according to this invention can be used to increase the specificity of hybridization of nucleic acids to the immobilized oligos. After hybridization, unhybridized strands can be washed away. Hybridized strands can then be released into solution without mixing. Released strands can be rebound to the immobilized oligos, and unhybridized strands can be washed away. Each successive release, rebinding, and washing increases the ratio of perfectly matched hybrids to mismatched hybrids.

An array can be "3'" or "5'". "3' arrays" possess free 3' termini and "5' arrays" possess free 5' termini. The immobilized oligos in a 3' array can be extended at their 3' termini by incubation with a nucleic acid polymerase. If it is a template-directed polymerase, only immobilized oligos hybridized to a template strand can be extended.

Methods of oligodeoxyribonucleotide synthesis directly on a solid support are also known in the art, including methods wherein synthesis occurs in the 3' to 5' direction (so that the oligos will possess free 5' termini). Methods wherein synthesis occurs in the 5' to 3' direction (so that the oligos will possess free 3' termini) are also known.

Suitable substrates or supports for arrays should be non-reactive with reagents to be used in processing, washable under stringent conditions, not interfere with hybridization and not be subject to inordinate non-specific binding. For example, treated glass polymers of various kinds (e.g., polyamide and polyacromorpholide), latex-coated substrates and silica chips.

Arrays can be made over a wide range of sizes. In the example of a square sheet, the length of a side can vary from a few millimeters to several meters.

II. Sorting nucleic acids

Our invention allows mixtures of strands to be sorted according either to their terminal oligo segments ("terminal sorting") or their internal oligo segments ("internal sorting") on a binary array.

There are two important aspects of our invention for sorting. First, each strand in a mixture can be made to hybridize at only a few, or a single, location. And second, each strand can be provided with universal terminal priming regions that enable PCR amplication without prior knowledge of the terminal nucleotide sequences and without the need to synthesize individual primers.

For terminal sorting, the priming region(s) can be made essentially dissimilar from the sequences occurring in the

nucleic acids that are present in the mixture to be sorted, so that priming does not occur anywhere but at the strands' termini. When strands from a complete restriction digest of a DNA are to be terminally sorted and amplified, priming only at the strands termini can be promoted by restoring the terminal restriction sites (those sites having been eliminated from internal regions by complete digestion) concomitant with the generation of terminal priming regions.

Terminal sorting is carried out on a binary array, which preferably is sectioned. The immobilized oligos contain a constant segment complementary to either the strands' 3' priming region or 5' priming region. Thus, each strand can only be hybridized to one location within the array. By sorting on a comprehensive array, every strand is bound somewhere within the array. This is especially important for the preparation of a comprehensive library of fragments of a long nucleic acid or a genome.

Strands can be sorted on either 3' or 5' arrays in which the constant segment is located either upstream or downstream of the variable segment. High specificity of sorting can be achieved by employing 3' arrays in which the constant segment of the immobilized oligos is upstream. In that case, sorting can be followed by the generation of an immobilized copy of each sorted strand using the immobilized oligos as primers for the synthesis of a complementary copy of that strand when the array is incubated with an appropriate DNA polymerase. The generation of copies covalently linked to the array enables the array to be vigorously washed to remove non-covalently bound material before strand amplification. It also enables the arrays to serve as permanent banks of sorted strands which can subsequently be amplified over and over to generate copies for further use.

A strand sorting procedure is shown in Figure 4. A DNA sample 10 is completely digested with a restriction endonuclease. The ends of each fragment are restored, and universal priming sequences 17 generated in the process to prepare fragments 11 for sorting. It is not necessary that priming sequences be added at both ends, if only linear amplification is desired. Nor is it

necessary that the priming sequence at the 3' end of a strand be the same as the priming sequence at the 5' end.

The strands are then melted apart 12 and hybridized to a terminal sequence binary sorting array, whose immobilized oligos 14 contain a variable segment 15 and a constant segment 16 which is complementary to the universal priming region 17, including the restored recognition site of the restriction enzyme 16a, 17a. Each strand is at a location dependent upon its variable sequence 100 adjacent to its priming sequence. At this point the array need not be sectioned. The array is then washed to remove unhybridized strands. The entire array is then incubated with DNA polymerase. Consequently, a complementary copy 18 of each hybridized DNA strand is generated by extension of the 3' end of the oligo to which the strand is bound. The array is then vigorously washed to remove the original DNA strands and all other material not covalently bound to the surface (not shown).

The covalently bound copy strands can be amplified. During amplification it is usually desirable that the array be sectioned. The wells are filled with a solution containing universal primers 19, 20, an appropriate DNA polymerase, and the substrates and buffer needed to carry out PCR. The array can, if desired, be sealed with a coversheet, further isolating the wells from each other. PCR is carried out simultaneously in each well of the array. This results in sorting the mixture of strands into groups of strands that share the same terminal oligo sequence, each strand (or each group of strands) being present in a different well of the array and amplified there.

The results of hybridization can be improved by "proofreading", or editing, the hybrids formed, by selectively destroying those hybrids that contain mismatches, without affecting perfect hybrids.

The length of the immobilized oligos in a strand sorting array is chosen to suit the number of strands to be sorted. When sorting strands according to their terminal sequences, the number of different strands obtained in each well equals the number of times that a particular oligo complementary to the variable segment of the immobilized oligo occurs among the termini of

different strands in the mixture. If the number of nucleotides in each variable segment is *n*, then the total number of such variable sequences is 4^n, and the mean number of different strands in a well is $N/4^n$, where N is the number of different strands in the mixture, provided that nucleotide sequence is random, and that each of the four nucleotides is present in equal proportion. If a random sequence that is the size of an entire diploid human genome (6×10^9 basepairs) is completely digested by a restriction endonuclease that has a hexameric recognition site, then the resulting mixture will contain approximately 3×10^6 strands with an average length of 4,096 nucleotides. If this mixture is then applied to a comprehensive binary array having variable segments eight nucleotides long, then each well will contain, on average, approximately 45 different strands.

Our invention also includes methods for isolating individual strands by sorting them according to the identity of their terminal sequences on sectioned binary arrays. The strands can be from restriction fragments or not, so long as unique priming sequences are added to at least one of the strand's termini, such as by methods described herein. If the number of different strands in a sample is rather small, there is a high probability that after the first stage of sorting, many wells will either not be occupied, or be occupied by only one type of fragment. In the case of a complex mixture of strands (such as from the digestion of an entire human genome), a number of different types of fragments will occupy each well. In that case, the isolation of individual fragments can be achieved by PCR amplifying the strands in each well in the first stage of sorting and then sorting the group of fragments from each well on a fresh sectioned array. After symmetric PCR amplification, each well of the first array will contain copies of the strands that were originally hybridized there, and also their complementary copies.

If the original strands were sorted by their 3' ends, then their copies in a given well will all possess the same 3'-terminal sequence, and their complementary copies will possess the same 5' end. However, the 3'-terminal sequences of the complementary copies of the original strands in each well will be

different (as will be the 5' terminal sequences of the original copies). Therefore, the complementary strands will bind at different locations within the new sectioned array, according to the identity of their own 3'-terminal sequences, and with a high probability, each of them will occupy a separate well, where they can then be amplified.

Alternatively, the second stage of sorting can be carried out according to the identity of the terminal sequences at the other end of each strand. For example, if the strands were sorted in the first stage by their 3' ends (on an array whose immobilized oligos contain upstream constant segments, then the groups of strands from each well in the first array can be sorted in a second stage by their 5' termini (on an array having downstream constant segments). In either procedure, as a result of the second round of sorting, almost all of the different types of fragments are separated from one another (with the exception of virtually identical allelic strands from a diploid genome, which usually have identical termini, and consequently are sorted into the same well). The isolated strands can then be used for any purpose. For example, they can be inserted into vectors and cloned, or they can be amplified and their sequences determined.

Our invention also includes the use of binary arrays for isolating selected strands by sorting according to the identity of terminal sequences. Strands can, for example, be selected that contain particular regions (such as genes) of special interest from a clinical viewpoint. After the relevant portion of a genome has been sequenced, an array can be made using only preselected oligos whose variable segments uniquely match the terminal sequences of the strands of interest, i.e., they would be long enough to uniquely hybridize to the desired strands.

Our invention also encompasses methods that include sorting fragments according to their internal sequences. When so sorting, strands may bind at more than one well. This type of sorting can be useful for a number of applications, such as the isolation of strands that contain particular internal sequence segments (utilizing a sectioned ordinary array), or the sorting of strands according to the identity of variable oligo segments

adjacent to internal restriction sites of a particular type (utilizing a sectioned binary array). The latter approach is useful for ordering sequenced restriction fragments. The sorting of strands by their internal segments on a 3' sectioned ordinary array is useful for the generation of partial strands by virtue of extension of the immobilized oligos.

Our invention includes the sorting, in particular for sequencing, of natural mixtures of RNA molecules, such as cellular RNAs. Establishing messenger RNA sequences is useful, not only for the identification and localization of genes in the genomic DNA, but also for providing information necessary to determine the coding gene sequences (i.e. the exon/intron structure of each gene). Furthermore, the analysis of cellular RNAs in different tissues, at different stages of development, and in the course of a disease, will clarify which genes are active. Usually, RNAs are short enough to be sorted and analyzed without preliminary fragmentation.

III. Preparing partial strands of nucleic acids on sectioned arrays

Our invention includes methods of using sectioned arrays for preparing all possible partial copies of a strand or a group of strands. Preparing complete sets of partials of a strand(s), and sorting the partials by their variable ends is especially useful in a process for determining the sequence of the strand or strands. The preparation of partials is accomplished by either of the following methods: (1) terminally sorting on sectioned binary arrays a mixture of partial strands generated by degradation of a "parental" strand(s) at random; or (2) generating partials on a sectioned ordinary array, through the sorting of a parental strand(s) according to the identity of the strand's internal sequences, followed by the synthesis of (complementary) partial copies of the parental strand(s) by the enzymatic extension of the immobilized oligos, utilizing the hybridized parental strands as templates, and then copying the immobilized partials.

By using comprehensive arrays, it is possible to prepare every possible one-sided partial of a strand.

In the first case (partialing before sorting), a strand, or a double-stranded fragment, or a group of either, carrying terminal priming regions, (these can be a strand or a group of strands sorted on a sectioned binary array as described above), is randomly degraded by a chemical or an enzymatic method, or by a combination of both. Then the mixture of partials is sorted on a sectioned binary array according to the identity of their newly generated termini, essentially as described above for the sorting of full-length strands by their terminal sequences, with new priming sites being introduced at these new termini either before or after sorting. Only those partials that possess both the newly introduced priming site and the already existing priming site (at the opposite end), will be amplified by subsequent PCR. Partials can be sorted according to the identity of a variable sequence at either their 3' termini or their 5' termini. However, as is the case for the sorting of full-length strands, the highest specificity can be achieved by sorting according to the identity of a variable sequence at the 3' termini, and carrying out the sorting on 3' arrays having upstream constant segments, or by sorting according to the identity of a variable sequence at the 5' termini, and carrying out the sorting on 5' arrays having downstream constant segments. In these cases, sorting can be followed by the generation of immobilized (complementary) copies of the sorted partials. The arrays with the immobilized copies can serve as permanent banks of the sorted partials which can subsequently be amplified over and over to generate copies for further use. Following sorting, each well in the array will contain immobilized copies of all of those partials whose variable end is complementary to the variable segment of the immobilized oligo. The other (fixed) end of these partials will be identical to one of the ends of the parental strands. If an oligo segment occurs more than once in a strand, or if it occurs in more than one strand in the group of strands subjected to partialing, then the well will contain a

corresponding number of different partials, all sharing the same sequence at their variable ends.

In the second case (sorting before partialing), partials are prepared directly from the parental strands that are hybridized to a sectioned ordinary array without prior degradation. A strand, or a mixture of strands, is hybridized to a 3' ordinary array. The immobilized oligos are then used as primers for copying the hybridized strands, beginning at the location within each bound strand where hybridization occurred, and ending at the upstream terminus of each bound strand. After extension of the immobilized oligos, the hybridized parental strands are discarded. At this point the wells contain immobilized (complementary) partial strands. The partials in one well all share a 5'-terminal oligo segment that is complementary to a particular internal oligo in the parental strand(s). The partial strands have 3'-terminal sequences that include the complement of the 5'-terminal region of the parental strand(s) (which contains a priming region). Unlike the methods described above for partialing before sorting, the immobilized complementary partials will contain a priming region at only one end and therefore can not be amplified exponentially. However, their linear amplification is possible, with the partials being synthesized as DNAs or RNAs. Where RNA partials are generated, the priming region at the partial copy's 3' terminus contains an RNA polymerase promoter. Synthesis of RNA copies is more efficient than linear synthesis of DNA copies. Alternatively, the synthesized copies can be provided with second priming regions and can then be amplified in an exponential manner by PCR. This approach is illustrated, schematically, in Figure 5.

Figure 5 illustrates the generation of partials for one DNA parental strand 30 on a 3' sectioned ordinary array. First, the strand 30 (many copies, of course) such as obtained from well 13a of sorting array 13, is hybridized to the partialing array 31, a 3' sectioned ordinary array, containing well 31a. The parental strand 30 binds to many different locations within the array, dependent on which oligo segments are present in the strand. A hybrid 32 is formed in each well at the array that contains an

immobilized oligo complementary to a strand's oligo segment. After hybridization, the entire array is washed and incubated with an appropriate DNA polymerase in order to extend the immobilized oligos utilizing the hybridized strand as a template. Each extension product 33 strand is a partial (complementary) copy of the parental strand. Each partial begins at the place 32 in the strand where hybridization occurred and ends at the strand's terminus. The strand preferably terminates at its 5' terminus with a universal priming sequence 17, such as one introduced into all strands when sorting strands on a sectioned binary array as described. This allows for amplification of the partials. That priming sequence can contain a restored restriction site 16a. The parental strand may also contain, if it was previously sorted on a binary sorting array, a priming sequence at its 3' terminus 17, adjacent to the variable sequence 100 that the strand was previously sorted by.

The entire array is then vigorously washed under conditions that remove the parental DNA strands and other material, preferably all, that is not covalently bound to the surface. The areas of the array then contain immobilized strands 33 that are complementary to a portion of the parental strand. The wells can then be filled with a solution containing the universal primer (or promoter complement), an appropriate polymerase, and the substrates and buffer needed to carry out multiple rounds of copying of the immobilized partial strands. The array can then be sealed, isolating the wells from each other, and (linear) copying can be carried out simultaneously in all of the wells in the array.

IV. Surveying oligonucleotides with binary arrays

Our invention includes using binary arrays to survey oligos contained in strands and partials. Binary arrays allow surveying to be improved as compared with ordinary arrays, and they allow new types of selective surveying (such as surveying "signature oligonucleotides").

In surveying, strands first can be randomly degraded into pieces whose average length slightly exceeds the surveyed length. After degradation, each resulting nucleic acid piece is ligated to the same type of oligo (i.e., a constant sequence), that preferably does not occur anywhere in the internal regions of the pieces. For example, the sequence of the added oligo can contain the recognition site of a restriction endonuclease that was used to digest the DNA prior to fragment sorting. The ligation can be carried out in solution prior to hybridization, or after hybridization of the pieces to binary immobilized oligos whose constant segment is complementary to the oligo to be ligated. Preferably, a 3' array is used, having upstream constant segments. The immobilized oligos can then be extended with an appropriate DNA polymerase, using the hybridized nucleic acid pieces as templates. It is preferable that after extension all hybrids have the same length. This can be achieved by employing dideoxynucleotides as substrates for the polymerase, to restrict extension to one nucleotide.

Hybrids can be labeled in both a ligation-dependent and an extension-dependent manner to increase the specificity of hybrid detection. Also, the ligated oligos and the added dideoxynucleotides can be tagged with different labels, for example, fluorescent dyes of different colors. The array is then scanned at two different wavelengths, and only those areas that emit fluorescence of both colors indicate perfect hybrids.

Survey results can be improved further by hybrid proofreading, by destroying hybrids containing mismatches, and by using chemical or enzymatic methods.

V. Use of the oligonucleotide arrays for the sequencing of nucleic acids

The arrays and methods of this invention can be used to determine the nucleotide sequence of nucleic acids, including the sequence of an entire genome, whether it is haploid or diploid. This embodiment requires neither cloning of fragments nor preliminary mapping of chromosomes. It is especially significant that

our method avoids cloning, a labor-intensive and time-consuming approach that is essentially a random search for fragments. In a preferred embodiment a comprehensive collection of whole nucleic acids or fragments is sorted into discrete groups. The sorted nucleic acids are then amplified with a polymerase, preferably by PCR.

Sequencing large diploid genomes, such as a human genome, using the arrays and methods of this invention is shown in Figure 6. We will describe the overall method in general terms. In the embodiment illustrated in Figure 6 an individual's genomic DNA 40 is digested with a restriction endonuclease and sorted by terminal sequences into groups of strands using a 3' sectioned binary sorting array 13, as is described above in Section II and illustrated in Figure 4.

Next, treating each well 13a of the sorting array separately, a complete set of partials is prepared for each group of sorted strands using a sectioned array 31, as is described above in Section III and illustrated in Figure 5. The partials can be generated in any chosen manner to make them detectable.

Then the contents of each well 31a of the partialing array 31 is surveyed using a survey array 42, as is described above in Section IV. Preferably the survey array is a binary array, but an ordinary array may be used. In the embodiment shown in Figure 6, surveying is performed with a sheet 43 containing miniature survey arrays 42 that have been printed in a pattern that coincides with the number and location of the wells 31a. The oligo information obtained can be used, according to our invention, to separately determine the nucleotide sequence of every strand in each group isolated on the sorting array.

To determine the order of the fragments sequenced as illustrated in the embodiment of Figure 6, genomic DNA 40 is digested with at least a second restriction endonuclease and sorted into groups of strands using a 3' sectioned binary sorting array 44, as is described above in Section II and illustrated in Figure 4. The contents of each well 44a of the sorting array 44 is surveyed with special survey arrays 45, 46 that identify "signature oligonucleotides" (described below) in intersite

segments of sorted fragments from different digests. This is done to determine the order of the fragments relative to one another without regard to differences between allelic pairs of fragments. In the embodiment shown in Figure 6 this surveying is performed with printed sheets 47, 48 that have been printed with a pattern of miniature arrays 45, 46.

To allocate the ordered allelic fragments to their respective chromosomes in a diploid organism, fragments are linked according to their allelic differences. In the embodiment illustrated in Figure 6, the strands from selected wells of the sorting array 44 are transferred to a selected well of one of a series of partialing arrays 49, partials are generated, and the partials are surveyed using miniature survey arrays 50 on printed sheets 51. Only the presence of oligos containing allelic differences in the selected partials needs to be determined to link a pair of allelic fragments to their respective neighboring allelic fragments.

When sorting according to the identity of terminal sequences, each strand occupies a particular "address" in the array. It is convenient to think of the address as the oligo sequence within a strand that directs the DNA strand to hybridize to a particular location, i.e., the sequence that is perfectly complementary to the variable sequence of the oligo immobilized at that location. The "address" also identifies the location within the array where the DNA binds.

After sorting, each group of strands is amplified and subjected to partialing. Importantly, the isolation of individual strands is not necessary, because our method allows the nucleotide sequence of each strand in a mixture to be determined. In particular, our method allows the sequences of strands in a well of the sorting array to be determined, separately from mixtures of strands in other wells. In a preferred embodiment, the partialing array is comprehensive in order to obtain all possible one-sided partials (i.e., a comprehensive array). Each group of partials is amplified prior to surveying. Most preferably, the amplification is carried out in such a manner that one

of the two complementary partial strands is produced in great excess over the other.

Each group of partials is surveyed to identify their constituent oligos. Surveying is preferably carried out using binary arrays.

Although not necessary, it is preferable to have the survey arrays be as compact as possible. It is anticipated that surveying will be advantageously accomplished simultaneously for many or all wells of a partialing array by utilizing a sheet on which miniature survey arrays have been "printed" in a pattern that coincides with the arrangement of wells in the partialing array, in a manner similar to that shown in Figures 6 and 7. Referring to Figure 7, partialing array 31, comprising an array of wells 31a, is surveyed using sheet 43, having printed thereon an array of miniaturized survey arrays 42. The pattern of arrays 42 corresponds to the pattern of wells 31a, whereby all wells 31a can be surveyed simultaneously.

Automated photolithography techniques for preparing miniature oligo arrays have been developed [Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. and Solas, D. (1991). Light-Directed, Spatially Addressable Parallel Chemical Synthesis, *Science* 251, 767-773]. The manufacture of miniature arrays on a "chip", for use in surveys also has been reported.

Surveying with comprehensive arrays produces a complete list of oligos contained in the partials in each well of the partialing array. This will reveal all oligos present in all partials in that well. The method of this invention can determine the sequences of the original (parental) fragment strands.

The "partials" referred to in this section are one-sided partial strands that begin at the 5' terminus of a parental nucleic acid strand (the fixed end) and end at different nucleotide positions in the strand (the variable end). Partials are sorted in the partialing array according to the identity of their variable ends, and therefore each partial has a particular "address" within the array. As with sorting arrays, an "address" in a partialing array is the oligo sequence that is present at the variable end of the partial strand and that is complementary

to the variable segment of an immobilized oligo. The "address" also relates to the location within the array where the partial strand is found, since the variable segment of the oligo immobilized in that well is complementary to the oligo at the partial's variable terminus. The "address" also relates to the location within the parental strand of a partial's terminal oligo. The location of this "address oligo" within a parental strand is characterized by an "upstream subset" of oligos that come before it in the parental sequence and by a "downstream subset" of oligos that come after it.

Our method of establishing nucleic acid sequences, for either a single strand or a group of parental strands sorted by their terminal sequences, begins by assembling an "address set" for each address in the partialing array. The "address set" is a comprehensive list of all oligos in all the parental strands which have the address oligo within their nucleotide sequences. The "upstream subset" contains all the oligos that occur upstream (i.e., towards the 5' end) of the address oligo in parental strands that contain the address oligo. The "downstream subset" contains all the oligos that occur downstream (i.e., towards the 3' end) of the address oligo in any parental strands that contain the address oligo. Together the two subsets form the "address set."

The upstream subset of each address can be determined directly from the survey of each well of a partialing array and consists of a list of all the oligos identified as being present in the partial strands in that well. The downstream subset of each address can be inferred by examining the upstream subsets of all the addresses: the downstream subset of a particular address consists of those addresses whose own upstream subset includes that particular address oligo.

The upstream subset and the downstream subset of a particular address, taken together, are an "indexed address set". If an oligo occurs more than once in a strand, it can occur in both the upstream and the downstream subsets of an address. Indexed address sets provide the information required to order the oligos contained in a strand set, as will be described below.

When a mixture of strands is examined, it is also useful to consider an address set without regard to which oligos occur upstream and downstream of an address. This is called an "unindexed address set". Unindexed address sets are decomposable into strand sets by the method of this invention.

We have discovered that when assembling big strand sets whose oligos do not all overlap uniquely, it is advantageous to work with "sequence blocks" rather than with individual oligos. Sequence blocks are composed of oligos that uniquely overlap one another in a given strand set. Two oligos contained in a strand set are said to overlap if they share a terminal (5' or 3') n-1 nucleotide sequence. An overlap is unique if no other oligo than those two in the strand set has this sequence at its termini. Here n is the length (in nucleotides) of each of the two oligos if they are of the same length or, if they are of different length, n is the length of the shorter one. We use unique overlaps to construct sequence blocks from the oligos in a strand set.

The position of each sequence block relative to the others is determined from the distribution of the oligos between the upstream and downstream subsets of every address. This is accomplished by finding, for each of the blocks, which blocks occur upstream, and which blocks occur downstream, of that block by examining the address sets. The address sets are used in order to generate "block sets." The block sets are address sets wherein blocks have been substituted for the oligos that comprise the blocks, including the address oligo. Once the relative position of the sequence blocks has been determined, they can be assembled into the final sequence. The assembly is governed by the following rules: (1) each of the blocks must be used at least once, (2) the blocks must be assembled into a single sequence, (3) the ends of neighboring blocks must match each other (i.e., overlap by an n-1 nucleotide sequence, see above) and (4) the order of the blocks must be consistent with their positions relative to one another, as ascertained from the block sets, as will be clear from the examples.

A sequence block can occur either once in a sequence, or more than once, and this we determine by examining the block sets. If a block occurs more than once in a sequence, it will always be contained in both its own upstream and downstream subsets. On the other hand, if a block occurs only once in a sequence, it may or may not be present in its own upstream or downstream subset. But, if a block is absent from either its upstream subset, or from its downstream set, that block occurs in the strand only once. The relative order of these "unique" blocks can be determined by noting which of them occur in the upstream subset, and which of them occur in the downstream subset, of the others. Once the unique blocks have been ordered relative to each other, the gaps between them are filled with blocks that may be non-unique. However, not every gap can necessarily be filled in with a particular block. There is a range of locations within which each non-unique block (or presumably-non-unique block) can be present. The range for a particular block is determined by noting those blocks that always occur upstream of it, and those blocks that always occur downstream of it. A gap can be filled in if, and only if, there is a block or a combination of blocks, whose outer ends have n-1 nucleotide-long perfect sequence overlaps with the ends of the blocks that form the gap. Because at least two overlaps, each of low probability, must occur simultaneously, it is highly unlikely that more than one block, or one combination of blocks, can fill a gap. If a particular block occurs many times in a strand, it will have to be used to fill every gap it matches. This is why, using the method of the invention, it is possible to establish the sequence of a strand without measuring how many times an oligo occurs in the partials. It is only necessary to determine whether an oligo is present or not.

An important aspect of this invention is the ability to sequence a mixture of strands simultaneously. The invention can be used for the determination of fragment sequences from an entire fragmented and sorted genome.

If one strand is being sequenced, all address sets determined from a partialing array will contain the same oligos that

constitute the strand set. The only difference is that some oligos which are downstream in one set may be upstream in another address set. If a mixture of strands have been partialed on a single partialing array, certain addresses will be shared by more than one parental strand. Their address sets will be composite, containing all of the oligos from all of the strands that the address oligo is present in. Addresses that are only found in a particular strand in the mixture, however, will have address sets which only contain oligos from that strand. They are identical to the strand set, and each contain the same oligos. The mixture can contain up to a hundred or so different DNA strands, each of a different length and sequence, as can be obtained with an appropriate sorting array (or set of sorting arrays) and method described above. When a mixture of strands is analyzed on a partialing array, the data obtained by surveying the partials will reflect the diversity of the sequences in the mixture, and will appear to be very complex. However, we have discovered a way to decompose the unindexed address sets obtained by analysis of a strand mixture into their constituent strand sets. Then, as we have described for sequencing a single strand, the oligos in each of the identified strand sets can be grouped into sequence blocks that can be ordered from the information contained in the indexed address sets, as will be clear from the examples.

Unindexed address sets can be either "prime" or "composite." A prime set consists of one strand set; while a composite set consists of more than one. A prime set cannot be decomposed into other address sets, i.e., there is no address set which is a subset of a prime set. Composite sets, however, can usually be decomposed into two or more simpler address sets. Once individual strand sets have been identified, they can each be treated as though they were obtained from an analysis of a homogeneous strand. It is thus possible, in many cases, to sequence all strands in an unknown heterogenous DNA sample without first isolating the strands.

The fragment sequences obtained by the methods outlined above or by any other method can then be put in their correct order using oligo arrays. Assembling restriction fragments into

contiguous sequences can be accomplished by identifying each fragment's immediate neighbors. One method for obtaining this information is to use another restriction enzyme to cleave the same DNA at different positions, thus producing a set of fragments that partially overlap neighboring fragments from the first digest, and then to sequence these fragments. However, it is not necessary to sequence the fragments in the second restriction digest. It is only necessary to uniquely identify overlapping segments in the fragments from alternate restriction digests. This can be done by surveying "signatures".

Signatures can be determined by hybridization of fragment strands to complementary oligo probes. A signature of a fragment may consist of one, two or more oligos, so long as it is unique within the sequence analyzed. Neighboring fragments from one restriction digest can be determined by looking for their signatures in overlapping fragments from an alternate digest.

We have devised a method for identifying neighboring restriction fragments among the list of sequenced fragments that does not require either cloning or sequencing of overlapping fragments. If strands from an alternate digest are sorted, complementary strands of the same fragment will hybridize to different addresses in the sorting array. Whenever intersite segments from two or more fragments of the first digest are present within one fragment of the second digest, then all of these segments will be represented in both complementary strands of that one fragment, and all will be present wherever those strands bind in a sorting array. We identify the segments by obtaining their signatures through hybridization to specialized binary survey arrays. The signatures of intersite segments that occur in one fragment always accompany each other, whereas signatures of distant segments travel independently.

After the fragments from an original (first) restriction digest of a long DNA have been sequenced, the same DNA is digested with a second (different) restriction endonuclease, the termini of the generated fragments are provided with universal priming regions (that also restore the recognition sites at the termini), and the strands are sorted according to particular

internal sequences, namely, a variable sequence adjacent to the recognition site for the first restriction enzyme. The sorting array is a sectioned binary array. It contains immobilized oligos having a variable sequence as well as an adjacent constant sequence that is complementary to the recognition sequence of the first restriction endonuclease. The sorted strands are amplified by "symmetric" PCR, so that in each well where a strand has been bound, copies of the bound strand, as well as complements, are generated. In another embodiment, strands can be sorted according to their terminal sequences on an array whose oligos' constant segments include sequences that are complementary to the recognition site of the second restriction enzyme. This alternative is not detailed, but it corresponds to the embodiment discussed below, but with terminal sorting.

Each strand that hybridizes to the binary sorting array will possess at least two recognition sites for the second restriction enzyme (restored at the strand's termini), and at least one (internal) recognition site for the first restriction enzyme. The segments included between these two types of restriction sites (intersite segments) comprise the overlaps between the two types of restriction fragments, and each intersite segment is thus bounded by any two restriction sites of the two types. It follows, that each of these segments can be characterized by identifying these two restriction sites and variable sequences of preselected length within the segment that are immediately adjacent to each of the restriction sites. The combination of a recognition site (for either the first or the second restriction enzyme) and its adjacent variable oligo we call a "signature oligonucleotide". Every intersite segment can be characterized by two signature oligos (of either type) that bound that segment. The combination of the two signature oligos is defined herein as the intersite segment's "signature".

After strand amplification, the strands in the wells of the sorting array are surveyed to identify the signature oligos of each of the two types. This is carried out by using two types of binary survey arrays. The first has immobilized oligos containing a variable oligo segment and a constant segment that is, or

includes, an adjacent sequence that is complementary to the recognition site for the first restriction endonuclease. The immobilized oligos in the second survey array has a variable oligo segment of preferably the same length as the variable segment of the first specialized survey array, and a constant segment that is, or includes an adjacent sequence that is complementary to the recognition site for the second restriction endonuclease. The constant oligo segments in these arrays can be located either upstream or downstream of the variable oligo segments, resulting in the surveying of either the downstream or the upstream signature oligos in each strand of the intersite segments being surveyed. In a preferred embodiment the constant oligo segments are upstream, and the immobilized oligos have free 3' ends, so that they can be extended by incubation with a DNA polymerase. From the oligo information that is obtained, the sequenced fragments can be ordered relative to one another.

In our method, the uniqueness of a signature is achieved by surveying "half signatures" (signature oligonucleotides) on two relatively small survey arrays. If the variable segments in the arrays are 8-nucleotide-long, the number of areas in the two arrays is approximately 130,000, or approximately 100,000,000 times smaller than the single array that would be needed for detecting the same size signature (28 nucleotides).

If a diploid genome (such as a human genome) is sequenced, the ordered fragments will appear as a string of unlinked pairs of allelic fragments. What remains unknown is how the allelic fragments in each pair are distributed between the homologous (sister) chromosomes that came from each parent. Allocation of the allelic fragments to these "chromosomal linkage groups" requires knowledge of which fragment in each pair is linked to which fragment in a neighboring pair.

We have developed a method that uses arrays for allocating allelic fragments to chromosomes, irrespective of what method was used for sequencing and ordering the fragments. The linkage of fragments in neighboring pairs can be achieved by sequencing a restriction fragment ("spanning fragment") from an alternate digest that spans at least one allelic difference in each pair.

Since the sequences of the allelic fragments are known, there is no need to sequence the spanning fragment. Instead, one can simply determine which oligos that harbor allelic differences accompany one another in the spanning fragment, i.e., which oligos occur in the same chromosome. This can be accomplished by surveying, at a selected address in a partialing array, partials generated from a selected group of restriction fragments from an alternate digest. A group of restriction fragments is selected that contains a spanning fragment, and an address in a partialing array is selected that encompasses a difference in one of the neighboring allelic pairs.

Since the sequence of every fragment is known, it is possible to choose an alternate restriction fragment that spans the allelic differences in the neighboring pairs. A spanning restriction fragment, in fact, may already be present at a particular address in one of the sorting arrays used to sort alternate digests during the ordering procedure.

In this method, sorted strands are melted apart, and the mixture is hybridized to a particular well in the partialing array, whose address corresponds to one of the allelic oligos. Two different wells are selected, each with an address that corresponds to an oligo that harbors a differenct allelic oligonucleotide After amplification of the partial strands, the oligos in the two wells are identified with a survey array. Examination tells which fragments are on the same chromosome.

Since allelic differences occur roughly once every 1,000 basepairs in the human genome, most allelic fragments resulting from digestion with a restriction enzyme recognizing a hexameric sequence (resulting in about 4,096 average length) will differ from each other. If the variable oligo segments in the survey arrays are made of octanucleotides, then each allelic nucleotide substitution will give rise to eight different oligos in each of the allelic fragments. However, using our method, inspection of only one address in the partialing array is sufficient to reveal the linkage of the corresponding reference oligo to any one of the eight oligos that encompass the nucleotide substitution that occurs in the neighboring fragment on the same chromosome.

Therefore, only one address in the partialing array is needed to reveal the linkages between two neighboring allelic pairs. Thus, 65,536 linkages can be determined on a single comprehensive partialing array made of variable octanucleotides. With this method, only 10 to 20 of these arrays would be needed to complete the assembly of an entire diploid human genome that has been fragmented by a restriction endonuclease with a hexameric recognition site.

Computational methods can be developed to minimize or eliminate errors that occur during partialing and surveying, by taking advantage of the high redundancy in the data. Such methods should take into account the following aspects of a preferred sequencing procedure: the sequence of every fragment is independently determined four times (by virtue of each strand and its complement being present at two different addresses in the sorting array); each strand set is determined in as many trials as the number of different oligos in that strand; every nucleotide in a strand is represented by as many different oligos as the length (of the variable segment) of the immobilized oligos in the survey array; the locations where a particular block can occur in a sequence are limited by the distribution of the blocks among the upstream and downstream subsets of each pertinent address; and the edges of a block must be compatible with the edges of each gap where that block is inserted.

Using our genome sequencing method, one can use throughout essentially the same technology, i.e., hybridization of oligo probes and the amplification of nucleic acids by the polymerase chain reaction, both of which are well-studied, common laboratory techniques. The entire procedure can be performed by a specially designed machine, resulting in huge reductions in time and cost, and a marked improvement in the reliability of the data. Many arrays could be processed simultaneously on such a machine. The machine most preferably should be entirely computer-controlled, and the computer should constantly analyze intermediate results. As stated above, used arrays can be stored, both to serve as a permanent record of the results, and to provide additional

material for subsequent analysis or for manipulating the sequenced strands and partials.

Analysis of an individual's genomic DNA provides the complete nucleotide sequence of that individual's diploid genome. The genes and their control elements are allocated into chromosomal linkage groups as they appear in a single living organism. The sequence will describe an intact, functioning ensemble of genetic elements. This complete sequencing provides the ability to compare genomes of individuals, thereby enabling biologists to understand how genes function together and to determine the basis of health and disease. The genomes of any species, whether haploid or diploid, can be sequenced.

The invention can be used not only for DNA's but as well for sequencing mixtures of cellular RNAs.

The invention is also useful to determine sequences in a clinical setting, such as for diagnosis of genetic conditions.

VI. Manipulating Nucleic Acids on Sectioned Arrays

Our invention also includes using sectioned arrays for introducing site-directed mutations into sequenced nucleic acids, including the introduction of nucleotide substitutions, deletions and insertions. This can be carried out in a massively parallel fashion. In one embodiment, a partial whose variable end has been deprived of a priming region, is ligated to the free terminus of an immobilized oligo that contains the mutation to be introduced. In another procedure, where the purpose of mutagenesis is to introduce a single-nucleotide substitution, then the substituting nucleotide can be added directly to the variable end of the partial. In both cases, the modified partials or their complementary copies are used to synthesize a mutant strand utilizing as a template either the complementary parental strand (i.e., from which the partials were generated) or a longer complementary partial, or any other strand or partial that encodes the missing region. The fixed end of the mutant partial is provided with a priming region that is different from the corresponding priming region of the template strand. Therefore, only mutant strands are capable of subsequent amplification by

PCR. A single array can be used either to mutate many single positions in a gene, or to introduce mutations in many genes in one procedure.

Sectioned arrays can also be used for the massively parallel testing of the biological effects of the introduced mutations. For example, parallel coupled transcription-translation reactions can be carried out in the wells of a sectioned array following amplification of the mutant strands. It is thus possible to determine simultaneously, on the same sectioned array, the effects of many different amino acid substitutions on the structure and function of a protein.

VII. Examples

1. Sorting nucleic acids or their fragments on a binary oligonucleotide array whose immobilized oligos have free 3' termini, with constant upstream segments --

This method allows the immobilized oligos to serve as primers for copying bound strands, resulting in the formation of complementary copies covalently linked to the array.

1.1. Sorting restriction fragments according to their terminal sequences, following the introduction of terminal priming regions --

DNA is digested using a restriction endonuclease. Recognition sites for the restriction endonuclease are restored in solution by introducing terminal extensions (adaptors) that contain a sequence which, together with the restored restriction site, form a universal priming region at the 3' terminus of every strand in the digest. This priming region is later used for amplification by PCR. After melting fragments, the strands are sorted on a sectioned binary array. A sequence complementary to the generated priming region serves as both the constant segment of the immobilized oligos and as the primer for PCR amplification of the bound strands.

DNA to be analyzed is first digested substantially completely with a chosen restriction endonuclease, and the fragments

obtained are then ligated to synthetic double-stranded oligo adaptors. The adaptors have one end that is compatible with the fragment termini. The other end is not compatible with the fragments' termini. The adaptors can therefore be ligated to the fragments in only one orientation. The adaptors' strands are non-phosphorylated, which prevents their self-ligation. The strands in the restriction fragments have their 5' termini phosphorylated which results from their cleavage by a restriction endonuclease. This favors the ligation of the adaptors by a DNA ligase (such as the DNA ligase of T4 bacteriophage) to the restriction fragments, rather then to each other. Since DNA ligase catalyzes the formation of a phosphodiester bond between adjacent 3' hydroxyl and phosphorylated 5' termini in a double-stranded DNA, the phosphorylated 5' termini of the fragments are ligated to the adaptor strand whose 3' end is at the compatible side of the adaptor. The 3' termini of the fragments remain unligated. A DNA polymerase possessing a 5'-3' exonuclease activity (such as DNA polymerase I from *Escherichia coli* or Taq DNA polymerase from *Thermus aquaticus*) is then used to extend the 3' ends of the fragments, utilizing the ligated oligo as a template, concomitant with displacement of the unligated oligo. To make the ligated oligo resistant to the 5'-3' exonuclease, the ligated oligo can be synthesized from α-phosphorothioate precursors.

Although the oligo adaptors are provided in great excess during the ligation step, there is still a low probability that two restriction fragments will ligate to one another, rather then to the adaptor. To prevent this, the ligation products can again be treated with the restriction endonuclease used to generate the fragments, in order to cleave the formed interfragment dimers. The endonuclease will not cleave the ligated adaptors if they are synthesized from modified precursors (such as nucleotides containing N^6-methyl-deoxyadenosine), which are known and currently commercially available [e.g., from Pharmacia LKB]. Resistance of the ligated adaptors to digestion by the restriction endonuclease can be increased further if the ligated oligo is synthesized from phosphorothioates, and if phosphorothioate analogs of the nucleo-

side triphosphates are used as substrates for extension of the 3' termini.

After the priming regions have been added, the complementary strands are melted apart, such as by increasing temperature and/or by introducing denaturing agents such as guanidine isothiocyanate, urea, or formamide. The resulting strands are hybridized to a binary sorting array, such as by following a standard protocol for the hybridization of DNA to immobilized oligos. Hybridization is performed so that formation of only perfectly matched hybrids is promoted. The hybrids have a length which is equal to that of the immobilized oligos. The immobilized oligos are attached to the array at their 5' termini and contain constant restriction site segments adjacent to a variable segment of predetermined length. Each strand will be bound to the array at its 3' terminus. Its location within the array will be determined by the identity of the oligo segment that is located in the strand immediately upstream from the restored restriction site at its 3' end, and that is complementary to the variable segment of the immobilized oligo to which it is bound. After hybridization and washing away all unbound material, the entire array is incubated with a DNA polymerase, such as Taq DNA polymerase deoxyribonucleotide 5' triphosphates or the DNA polymerase of bacteriophage T7, and substrates. As a result, the 3' end of each immobilized oligo to which a strand is bound will be extended to produce a complementary copy of the bound strand. The array is vigorously washed. The wells are then filled with a solution containing universal primer, an appropriate DNA polymerase, and the substrates and buffer needed to carry out PCR. The array is then sealed, isolating the wells from each other, and exponential amplification is carried out, preferably simultaneously, in each well.

1.2. Sorting restriction fragments according to their terminal sequences, with 3' and 5' terminal priming regions being introduced, one before and one after strand sorting --

This procedure consumes larger amounts of enzymes and substrates than the procedure described in Example 1.1, however,

only those strands that are correctly bound to the immobilized oligos acquire both priming regions necessary for PCR. The possibility that non-specifically bound strands will be amplified is minimized. Furthermore, different priming regions can be introduced at different termini of a strand. It then becomes possible to: (1) perform "asymmetric" PCR, where only one of the complementary strands is accumulated in significant amounts, and remains single-stranded: (2) introduce a transcriptional promoter into only one of the priming regions, in order to be able to obtain RNA transcripts of only one strand (without also producing its complement; (3) differentially label complementary strands; and (4) avoid self-annealing of the strand's terminal segments that can interfere with primer hybridization and lower PCR efficiency.

In this example, digestion of DNA, adaptor ligation and re-digestion of fragments are carried out as described in Example 1.1, above. The 3' ends of the restriction fragments, however, are not extended by incubation with DNA polymerase. Instead, the strands ligated at their 5' ends to adaptors are melted apart from their unextended complements and hybridized to a binary array. The array contains immobilized oligos that are pre-hybridized with shorter complementary 5'-phosphorylated oligos that cover (mask) the immobilized oligos except for a segment which includes a variable region and a region complementary to the portion of the restriction site remaining at the fragments' (unrestored) 3' end. The masked region includes the rest of the restriction site and any other constant sequence, such as may be included in a priming region. Hybridization is carried out under conditions that promote the formation of only perfectly matched hybrids which are the length of the unmasked segment of the immobilized oligo. After washing away the unbound strands, the strands that remain bound are ligated to the masking oligos by incubation with DNA ligase. The correctly bound strands thus acquire a priming region at their 3' end, in addition to the priming region they already have at their 5' end. The two priming regions preferably correspond to different primers. The array is then washed under appropriately stringent conditions to

remove all nucleic acids except the immobilized oligos and the ligated strands hybridized to them.

1.3. Sorting RNAs according to their terminal sequences --

Mature eukaryotic mRNAs share structural features that can help in their manipulation using arrays. All have a "cap" structure on their 5' end, and most also possess a 3'-terminal poly(A) tail, which is attached posttranscriptionally by a poly(A) polymerase. Because there are usually no long oligo(A) tracts in the internal regions of cellular RNAs, the poly(A) tail can serve as a naturally occurring terminal priming sequence in sorting. The size of mRNAs (several thousand nucleotides in length) allows them to be amplified and analyzed directly, without prior cleavage into fragments.

There are known methods for preparing essentially undegraded total cellular RNA. Total cellular RNA is converted into complementary DNA (cDNA) using an oligo(dT) primer and a reverse transcriptase or *Thermus thermophilus* DNA polymerase. Then, omitting second strand synthesis, single-stranded cDNAs (which possess oligo(dT) extensions at their 5' end and variable 3' termini) are sorted according to their 3'-termini on a sectioned binary array and are ligated there to pre-hybridized adaptors of a predetermined sequence that are complementary to the immobilized oligos' constant sequence, and that introduce into a cDNA molecule the 3'-terminal priming site. The cDNA is amplified, using two primers for PCR: oligo(dT) and an oligo complementary to the adaptor.

2. Preparing partial strands of nucleic acids on oligonucleotide arrays --

There are two aspects to this procedure: first, the generation of partial strands (partials), and second, the sorting of partials according to their terminal oligo segments. All of the embodiments described below are based on the following principle: in generating partials from a strand, one of the original strand ends is preserved (it will be referred to as the "fixed" end), whereas the other end is truncated to a different extent in the

various partials (it will be referred to as the "variable" end). Although either the 5' or the 3' end of the original strand can serve as the fixed end, it is preferable that the 5' end be fixed. If amplification of sorted partials is desirable, it is preferable that the 5' end of the original strand, i.e., the fixed end, be provided with a priming region prior to partialing by any of the methods described above, and that partialing be carried out on a sectioned array. Either an individual strand or a mixture of strands can be subjected to a partialing; however, if the mixture is very complex (such as a restriction digest of a large genome), it is desirable that the mixture first be sorted into less complex groups of strands, as described above. The groups of strands used for preparing partials should essentially be devoid of contaminating strands; therefore, sorting by terminal sequences is preferable for the preliminary sorting. If preliminary sorting is performed, the strands will already contain terminal priming regions necessary for amplification of the partials. Partialing can be performed on either DNA or RNA, the final product being either DNA or RNA, in either a double-stranded or a single-stranded state.

2.1. Methods employing enzymatic cleavage of DNA fragments --

The purpose of the cleavage is to produce a set of partials of every possible length; therefore, DNA should be cleaved as randomly as possible, and to the extent that there is approximately one cut per strand. Deoxyribonuclease I (DNase I) cleaves both double-stranded and single-stranded DNA; however, double-stranded DNA is preferable as the starting material for preparing partials because of its essentially homogeneous secondary structure, so that every segment of a DNA molecule is equally accessible to cleavage. Double-stranded DNA fragments are produced as a result of "symmetric" PCR that can be carried out when sorting strands. An advantage of using DNase I is that it produces fragments with 5'-phosphoryl and 3'-hydroxyl termini, that are suitable for enzymatic ligation.

After cleavage of the double-stranded DNA fragments, DNase is removed, e.g., by phenol extraction. The (partial) strands are then melted apart and are hybridized to a sectioned binary array, wherein the immobilized oligos are pre-hybridized with shorter complementary 5'-phosphorylated oligos of a constant sequence that cover (mask) the immobilized oligos except for a segment that consists of a variable sequence. Hybridization is carried out under conditions that favor the formation of perfectly matched hybrids of a length that is equal to the length of the unmasked (variable) segment of the immobilized oligo, and that minimize the formation of imperfectly matched hybrids. After washing away unbound strands, the bound strands are ligated to the masking oligos by incubation with a DNA ligase. The ligated masking oligos will themselves serve as the second (3'-terminal) priming region of a partial strand. (All the partials of a strand will share the same 5' priming sequence that had been introduced into the strand before generation of the partials). If restriction fragments are to be partialed that possess some restriction site at their termini and do not possess this site internally, it is preferable that the 3' terminal priming region added to the partials include that site. This increases the specificity of terminal priming during subsequent amplification of the partials by PCR. Subsequent extension, washing, and amplification steps are as described in Example 1.1. If the partials are prepared for the purpose of sequence determination, asymmetric PCR can be performed. Alternatively, an RNA polymerase promoter sequence can be included in one of the two primers, and amplified DNA is then transcribed to produce multiple single-stranded RNA copies of one of the two complementary partial strands.

2.2. Methods employing chemical degradation of DNA --

These methods are applicable to both double-stranded and single-stranded nucleic acids. Chemical degradation is, in most cases, essentially random. It can be performed under conditions that destroy secondary structure, and the small size of the

modifying chemicals makes the chemicals readily accessible to nucleotides in secondary structures.

Both base-nonspecific reagents and base-specific reagents can be used. In the latter case, after base-specific cleavage is performed separately with several portions of the sample, the portions are mixed together to form a set of all possible partial DNA lengths. The main drawback to chemical cleavage is that the location of the terminal phosphate groups on the fragments is opposite to what is required for enzymatic ligation: 5'-hydroxyl and 3'-phosphoryl groups are produced in most cases. To overcome this problem, enzymatic dephosphorylation of 3' ends can be carried out.

2.3. Method of preparing partials directly on a sectioned array, without prior degradation of nucleic acids --

In this embodiment, the generation of partials and their sorting according to the identity of the sequences at their variable ends occur essentially in one step. First, a strand or a group of strands (if double-stranded nucleic acid is used as a starting material, the complementary strands are first melted apart), is directly hybridized to a sectioned ordinary array, whose oligos only comprise variable sequences of a pre-selected length, and that are immobilized by their 5' termini. Optimally, hybridization is carried out under conditions in which hybrids can only form whose length is equal to the length of the immobilized oligo. If the array is comprehensive, then a hybrid is formed somewhere within the array for every oligo that occurs in a DNA's sequence. After hybridization, the entire array is washed and incubated with an appropriate DNA polymerase in order to extend the immobilized oligo, using the hybridized strand as a template. Each product strand is a partial (complementary) copy of the hybridized strand. Each partial begins at the place in the strand's sequence where it has been bound to the immobilized oligo and ends at the priming region at the 5' terminus of the strand. If a priming region has not been introduced at the strand's 5' end before partialing, it can be generated at this step, after the hybrids that have not been extended, are elimi-

nated by washing. This can be done either by ligating the 5' end of the bound strand to a single-stranded oligoribonucleotide adaptor, or by tailing the immobilized partial copy with a homopolynucleotide. The entire array is vigorously washed under conditions that remove the original full-length strands and essentially all other material not covalently bound. Subsequent amplification of the immobilized partials can be carried out in different ways, dependent on whether it is desired to use linear or exponential amplification.

Exponential copying results in the generation of partials and their complements. For a strand to be exponentially amplified by PCR, both of its termini should be provided with a priming region, preferably different priming regions. The immobilized (complementary) partial contains only one (3'-terminal) priming region, and a complementary copy produced by linear copying would also have only one priming region (on its 5' end). For RNA copies to have a priming region at their 5' ends, the immobilized partial should have been provided with an RNA polymerase promoter downstream of its 3' terminal priming region using the methods described herein. The second priming region that is needed for exponential amplification can be introduced at the 3' ends of the complementary copies as follows.

(a) The 3' termini of RNA copies can then be ligated to oligoribonucleotide or oligodeoxyribonucleotide adaptors which are phosphorylated at their 5' end and whose 3' end is blocked. Exponential PCR can be performed by utilizing the two primers that correspond to the two priming regions, and then incubating with Tth DNA polymerase.

(b) If the amplified copies are DNA, they can be transferred, such as by blotting, (after melting them free of the immobilized partial) onto a binary array that is a mirror copy of the first array in the arrangement of the variable segments of its immobilized oligos. The constant segments of this binary array are pre-hybridized to masking oligos whose ligation to the 3' termini of the transferred DNAs (by DNA ligase) results in generation of the second priming region to permit exponential PCR.

In methods (a) and (b), both priming regions preferably contain, when applicable, the recognition sequence of the restriction endonuclease that was used to digest the genomic DNA before full-length strand sorting, and which had thus been substantially eliminated from the strands' internal regions.

(c) If partials are surveyed only for oligos that occur in one complementary strand (such as detecting only parental oligos), either only one of the two different primers should be labeled, or the primers should be labeled differently. It is also possible to use labeled substrates during asymmetric PCR.

3. Surveying oligonucleotides with binary arrays --

Surveying oligo content can be carried out in the different embodiments of the invention by hybridization of strands (or partials) to an ordinary array, followed by detection of those hybridized. However, the signal-to-noise ratio is not high enough to always avoid ambiguous results. The most significant problem is inability to sufficiently discriminate against mismatched basepairs that occur at the ends of hybrids. That hampers analysis of complex sequences. The use of binary arrays helps to overcome this problem.

Binary arrays are also useful for surveying longer oligos than are easily surveyed on an ordinary array (e.g., signature oligos) without increasing the size over that of an ordinary array.

Immobilized oligos in a binary survey array can have either free 5' or 3' ends, and the constant segment can be either upstream or downstream. In most cases, it is preferable that the 3' ends of immobilized oligos be free, and that their constant segments be upstream.

Surveying can utilize sectioned arrays. However, the use of plain arrays is preferable because they are less expensive and more amenable to miniaturization. The following methods are based on the use of plain binary arrays and involve fragmentation of the strands or partials prior to surveying.

3.1. Comprehensive surveys of DNA strands --

Every oligo present in a strand or in a partial, or in a group of strands or partials, is surveyed. If a survey of partials is performed in order to establish nucleotide sequences, it is preferable that each partial be represented by the same sense copies. Thus, there should be only one of the complementary strands in a sample or the complementary strands should be differentiable, e.g., one strand should produce either no detectable signal or a weaker signal. This can be accomplished by amplifying the partials linearly or by the use of asymmetric PCR.

DNA strands (or partials) to be surveyed are preferably digested with nuclease S1 under conditions that destabilize DNA secondary structure. The digestion conditions are chosen so that the DNA pieces produced are as short as possible, but at the same time, most are at least one nucleotide longer than the variable segment of the oligos immobilized on the binary array. If the surveyed strands or partials have been previously sorted and amplified on a sectioned array, this degradation procedure can be performed simultaneously in each well of that array. Alternatively, if it is desired to store that array as a master for later use, the array can be replicated by blotting onto another sectioned array. The DNA is then amplified within the replica array by (asymmetric) PCR prior to digestion with nuclease S1.

After digestion, the nuclease is inactivated by, for example, heating to 100°C, and the DNA pieces are hybridized to an array whose immobilized oligos' constant segments are prehybridized to 5'-phosphorylated complementary masking oligos. Preferably, the constant segment contains a restriction site that has been eliminated from the internal regions of the strands prior to sorting and is long enough so that its hybrid with the masking oligo is preserved during subsequent procedures.

The array is incubated with DNA ligase to ligate the masking oligos to only those hybridized DNA strands (or partials) whose 3' terminal nucleotide is immediately adjacent to the 5' end of the masking oligo, and matches its counterpart in the immobilized oligo. DNA ligase is especially sensitive to mismatches at the junction site.

After all non-ligated DNA pieces have been washed away under much more stringent conditions that were used during hybridization, the immobilized oligos are extended by incubation with a DNA polymerase, preferably by only one nucleotide, using the protruding part of the ligated DNA piece as a template, and preferably using the chain-terminating 2',3'-dideoxynucleotides as substrates. Extension is only possible, if the 3'-terminal base of the immobilized oligo forms a perfect basepair with its counterpart in the hybridized DNA piece. The use of the dideoxynucleotides ensures that all hybrids are extended by exactly one nucleotide and that all are of the same length. The array is then washed under conditions sufficiently stringent to remove unextended hybrids.

3.2. Detection of hybrids --

Hybrids can be detected by a number of different means. Unlabeled hybrids can be detected by using surface plasmon resonance techniques, which currently can detect 10^8 to 10^9 hybrid molecules per square millimeter. Alternatively, hybrids can be conventionally labeled, such as with radioactive or fluorescent groups. Fluorescent labels are convenient.

To ensure the lowest level of background labeling, it is preferable to label hybrids in a manner such that its detection is dependent on the success of both a ligation and an extension step. This can be accomplished within the scheme of oligo surveying by labeling the masking oligos, and the 2',3'-dideoxynucleotides used for the extension with fluorescent dyes possessing different emission spectra. The array can then be scanned at different wavelengths, corresponding to the emission maxima of the two dyes, and only signals from those areas that emit fluorescence of both colors are taken as a positive result.

After hybrids are extended (concomitant with labeling) and edited, the array is thoroughly washed to remove unincorporated label, destroy unextended hybrids, and discriminate one more time against mismatched hybrids that might have remained. A preferred method is to wash the array at steadily increasing temperature, with the signal from each area being read at a pre-determined

time, when the conditions ensure the highest selectivity for the particular hybrid that forms in that area. Other conditions (such as denaturant and/or salt concentration) can also be controlled over time. The fluorescence pattern can be recorded at predetermined time intervals with a scanning microfluorometer, such as an epifluorescence microscope.

4. Determination of the nucleotide sequences of strands in a mixture when each strand possesses at least one oligo that does not occur in any other strand in the mixture --

Figures 8 to 11 depict the determination of the sequences of two mixed strands using the methods of the invention. The example demonstrates the power of the invention to identify all the oligos present in a strand (i.e., its strand set) when it possesses at least one oligo that does not occur in any other strand in the mixture. In particular, the example demonstrates: (a) how the data obtained by surveying the partial strands generated from a mixture of strands and sorted by their variable termini (i.e., the upstream subset of each address) and the inferred downstream subset of each address (which together form the indexed address sets) are used to construct the unindexed address sets; and (b) how the unindexed address sets are compared to each other to identify prime sets. The example also demonstrates how the oligos contained in a strand set are assembled into the sequence of the strand, even though the primary data is obtained from a mixture. In particular, the example demonstrates: (a) how oligos in a strand set are assembled into sequence blocks; (b) how the contents of the indexed address sets are filtered so that only information pertaining to the oligos in a particular strand set remains; (c) how this filtered data is re-expressed in terms of the sequence blocks that are contained in that particular strand; (d) how information in the resulting "block sets" is used to identify those blocks that definitely occur only once in the strand ("unique blocks") and to identify those that can potentially occur more than once; (e) how information in block sets of unique blocks is used to determine the relative order of the blocks that occur only once in the strand;

(f) how the information in the block sets limits the positions at which the other blocks can occur (relative to other blocks); and (g) how a consideration of the sequences at the ends of blocks, in combination with a consideration of the relative positions of the blocks, leads to the unambiguous determination of the complete sequence of the strand. This example also illustrates: (a) how oligos that occur more than once in a strand are identified and located within the sequence, even though the survey data contain no information as to the number of times a particular oligo occurs in a partial or a mixture of partials having the same terminal oligo; and (b) how the sequences of different strands in a mixture can be determined separately, despite the fact that many of the oligos occur in more than one strand.

Figure 8a shows the sequences of two short strands (parental strands) that are assumed to be present in a mixture (with no other strands). It is assumed that complete sets of partials have been generated from this mixture, and that each set of partials has been separately surveyed, with the partials sharing the same address oligo being surveyed together. For the purpose of illustrating the method of analyzing the data, it is assumed that the address oligos and the surveyed oligos are three nucleotides in length. In practice, longer oligos should be used. However, for illustration it is easier to comprehend an example based on trinucleotides. The same methods of analyzing the data apply when longer oligos are surveyed, when much longer strands are in the mixture, and when the mixture contains many more strands.

Figure 8b shows the upstream subsets determined by surveying and the downstream subsets inferred (i.e., Figure 8b shows indexed address sets). The address oligos (bold letters) are listed vertically in the center of the diagram. The oligos listed horizontally to the left of each address oligo are those oligos that were detected in a survey of the partials at that address (the upstream subset). The oligos listed horizontally to the right of each address oligo are those inferred from the upstream subsets to occur downstream of that address oligo (the downstream subset). For example, oligo "ACC" is contained in the

upstream subset of the address oligo "CCT". This means that oligo "CCT" occurs downstream of oligo "ACC" in at least one strand in the mixture. Therefore "CCT" is inferred to be in the downstream subset of address set "ACC". The remaining downstream oligos in all of the address sets are similarly inferred. Note that an address oligo is a member of its own upstream and downstream subsets.

After the indexed address sets of all addresses in the parental strands have been determined (as shown in Figure 8b), the information is organized into unindexed address sets (Figure 8c), having no division into downstream and upstream subsets, but merely listing, for each address oligo, those oligos that occur in either the upstream or downstream subset (or in both). In Figure 8c, the address oligos (bold letters) are listed vertically on the left side of the diagram. Note that the address oligo is a member of its own unindexed address set.

Unindexed address sets are grouped together according to the identity of the oligos they contain (Figure 8d). Unindexed address sets that contain an identical set of oligos are grouped together. It can be seen that three groups of address sets are formed in this example. The groups are identified by the Roman numerals (I, II, and III). The address oligos of each group (for example, CTA, GTC, and TCC in group II) always occur together in a strand and can occur together in more than one strand.

Each group of identical address sets is then compared to all other groups of identical address sets to see if its common address set appears to be a prime by seeing whether any other address set is a subset of it. For example, in Figure 8d, the address set common to group III is not a prime address set, because the address set common to group I is a subset of the address set common to group III. However, the address set common to group I and the address set common to group II appear to be prime address sets.

Each putative prime address set is then tested to see if it is a strand set by examining all the address sets that contain all of the oligos that are present in it. For example, in Figure 9a, all the address sets that contain all the oligos present in

the putative prime address set common to group I are listed together (namely the address sets contained in groups I and III). The address oligos are shown in bold letters on the left side of the diagram, and the groups are identified by Roman numerals. The address set common to group I is indeed a prime address set (and therefore it contains a single strand set) because a list of the eleven oligos that are found in every address set in the diagram (they are seen as full columns) is identical to the list of eleven addresses on the left side of the diagram. Similarly, Figure 8b shows why the address set common to group II is also a prime set. The twelve oligos common to every address set in the diagram are all found in the list of twelve addresses on the left side of the diagram. Had either of these putative prime address sets not turned out to be a prime set (by the criterion described above), then it would have been identified as a pseudo-prime address set, and further analysis would have been required to decompose it into its constituent strand sets.

Once the strand sets in a mixture have been identified, the oligos in each strand set can be assembled into the strand sequence in a series of steps, as illustrated in Figure 10 (which utilizes the strand set determined in Figure 9a).

First the oligos in the strand set are assembled into sequence blocks. A sequence block contains one or more uniquely overlapping oligos. Two oligos of length n, uniquely overlap each other if they share an identical sub-sequence that is n-1 nucleotides long and no other oligos in the same strand set share that sub-sequence. For example, for the strand set shown in Figure 10a, the oligos "CAT" and "ATG" share the sub-sequence "AT" which does not occur in other oligos. These two oligos therefore uniquely overlap to form the sequence block "CATG", as shown in Figure 10b. Similarly, oligo "TGG" uniquely overlaps oligo "GGT" by the common sub-sequence "GG", and oligo "GGT" also uniquely overlaps (on its other end) oligo "GTA" by the common sub-sequence "GT". Thus, the three oligos ("TGG", "GGT", and "GTA") can be maximally overlapped to form sequence block "TGGTA". In forming sequence blocks, the following rule is adhered to: two oligos can be included in the same block if they

are the only oligos in the strand set to possess their common sub-sequence. Thus, "ATG" does not uniquely overlap "TGG", because the strand set contains a third oligo, "TTG", that shares the common sub-sequence "TG". If, following these rules, an oligo does not uniquely overlap any other oligo, then a sequence block consists of only that oligo. For example, "TAA" forms its own block. Following the above rules, the eleven oligos that occur in strand set A can be assembled into four sequence blocks.

Second, the data contained in the indexed address sets shown in Figure 8b are filtered to remove extraneous information that does not pertain to strand set A. Figure 10c shows the resulting filtered address sets. All address sets whose address oligo is not one of the oligos in strand set A are eliminated. In addition, all oligos that are not members of strand set A are removed from the upstream and downstream subsets of the remaining address sets. The resulting filtered address sets are then grouped together according to the oligos that are contained in each block. For example, the filtered address sets for address oligos "CAT" and "ATG" have been grouped together in Figure 10c because these two oligos are contained in sequence block "CATG". In Figure 10c, the address oligos found in the same block are identified by rectangular boxes. In addition, oligos that occur in the same block are grouped together within each upstream and downstream subset.

Third, the filtered address sets are converted into block sets, as shown in Figure 10d. In a block set, the information from different address sets is combined. Instead of a different horizontal line for each filtered address set that pertains to a particular block, the information in all of the address sets that pertain to that particular block is combined into a single horizontal line. For example, in Figure 9c, five different filtered address sets pertain to sequence block "TACCTTG". In Figure 10d, these five lines are combined into a single line in which the address oligos are replaced by an "address block", shown as "TACCTTG" surrounded by a bold box. Similarly, the upstream oligos are replaced by upstream blocks, and the downstream oligos are replaced by downstream blocks. In substituting

sequence blocks for the upstream (or downstream) oligos that are contained in the filtered address sets for a given address block, the following rule is adhered to: a sequence block only occurs in the upstream subset (or in the downstream subset) of an address block, if every oligo that is contained in that address block occurs in the upstream (or in the downstream) subset of every filtered address set that pertains to that address block. For example, sequence block "CATG" occurs in the upstream subset of address block "TACCTTG" because oligos "CAT" and "ATG" occur in the upstream subset of address oligos "TAC", "ACC", "CCT", "CTT", and "TTG".

Often, a sequence block does not occur in its own upstream or downstream subset. For example, sequence block "CATG" does not occur in the upstream or downstream subset of its own block set (i.e., in block set "CATG"), because oligo "ATG" is not present in the upstream subset of address set "CAT" and oligo "CAT" is not present in the downstream subset of address set "ATG". When a sequence block does not occur in its own upstream or downstream subset, this indicates that that sequence block occurs only once in the nucleotide sequence of that strand. However, a sequence block may occur in both the upstream subset and in the downstream subset of its own block set. For example, sequence block "TGGTA" occurs in both the upstream subset and in the downstream subset of block set "TGGTA". When a sequence block does occur in its own upstream and downstream subsets, it indicates that the sequence block may, but not must, occur more than once in the sequence. The presence of more than one parental strand in the original mixture can introduce additional oligos into the filtered upstream and downstream subsets that can cause a block that actually occurs only once in a sequence to appear in both the upstream and downstream subsets of its own block set. However, further analysis of the data determines the multiplicity of each block in the strand (as described below), thus resolving these uncertainties. For convenience, block sets that pertain to blocks that definitely occur only once in the sequence are listed together. For example, in Figure 10d, block set "CATG" and block set "TACCTTG" are listed together.

Fourth, the position of each sequence block relative to the other sequence blocks is determined. An examination of the block sets that pertain to unique blocks (that definitely occur only once in the sequence of the strand) indicates their relative positions. For example, in Figure 10d, block set "CATG" indicates that unique sequence block "TACCTTG" occurs downstream of unique sequence block "CATG". This is confirmed by block set "TACCTTG", in which unique sequence block "CATG" occurs upstream of unique sequence block "TACCTTG". The relative position of the two unique sequence blocks is indicated in Figure 10e, where the top line to the left of the arrow shows "CATG" upstream (to the left) of "TACCTTG". The relative position of the sequence blocks that can potentially occur more than once in the nucleotide sequence of the strand is determined from their presence or absence in the upstream and downstream subsets of other sequence blocks. For example, sequence block "TAA" occurs in the downstream subset of block set "CATG" (and does not occur in the upstream subset of block set "CATG"). Furthermore, sequence block "TAA" also occurs in the downstream subset of block set "TACCTTG" (and not in its upstream subset). Therefore, sequence block "TAA" must occur downstream of both unique sequence blocks "CATG" and "TACCTTG". This is indicated in Figure 10e, where the bottom line to the left of the arrow shows "TAA" as occurring downstream of "CATG" and "TACCTTG". Furthermore, sequence block "TGGTA" occurs only in the downstream subset of block set "CATG". Therefore, it must occur downstream of "CATG" in the sequence. On the other hand, sequence block "TGGTA" occurs in both the upstream and downstream subsets of block set "TACCTTG". This indicates that "TGGTA" can potentially occur in the sequence at positions both upstream and downstream of unique sequence block "TACCTTG". Finally, "TGGTA" only occurs upstream of "TAA". This is indicated in Figure 10e, where the bottom line to the left of the arrow contains a bracket that shows the range of positions at which "TGGTA" can occur, relative to the positions of the other sequence blocks. At this point in the analysis, the diagram to the left of the arrow in Figure 9c contains all the information obtained that pertains to strand set A.

Finally, the sequence of the strand is ascertained by taking into account both the relative position of the sequence blocks, as shown in the diagram to the left of the arrow in Figure 10e, and the identity of the sequences at the ends of the sequence blocks. The object of this last step is to assemble the blocks into the final sequence. Four rules are followed: (a) each of the blocks must be used at least once; (b) the blocks must be assembled into a single sequence; (c) the ends of blocks that are to be joined must maximally overlap each other (i.e., if the surveyed oligos are *n* nucleotides in length, then two blocks maximally overlap each other if they share a terminal sub-sequence that is *n*-1 nucleotides in length); and (d) the order of the blocks must be consistent with their positions relative to one another, as ascertained from the block sets. For example, in Figure 10e, "CATG" is upstream of "TACCTTG". "CATG" cannot be joined directly to "TACCTTG", since these two sequence blocks do not possess maximally overlapping terminal sequences (two nucleotides in length). However, an examination of the permissible positions at which other sequence blocks can occur indicates that "TGGTA" can occur in the gap between "CATG" and "TACCTTG". The ends of these sequence blocks are then examined to see whether the gap can be bridged. "CATG" can be joined to "TGGTA" by maximally overlapping their shared terminal sub-sequence "TG". Furthermore "TGGTA" can be joined to "TACCTTG" by maximally overlapping their shared terminal sub-sequence "TA". Similarly, the gap that occurs downstream of "TACCTTG" can potentially be filled by both "TAA" and "TGGTA". "TAA" must be used, because it was not used at any other location. However, "TACCTTG" cannot be directly joined to "TAA". The solution is to join "TACCTTG" to "TGGTA", and then to join "TGGTA" to "TAA". Thus, the sequence of strand A (which is shown in Figure 10f) is unambiguously assembled by utilizing sequence block "TGGTA" twice (as summarized in the diagram to the right of the arrow in Figure 10e).

The same procedure is followed to determine the sequence of strand B (see Figure 11). In this example, there are three sequence blocks that do not occur in their own upstream or downstream subsets, and they therefore definitely occur only once

in the sequence of strand B (namely, sequence blocks "CTTG", "GTCC", and "TACC"). An examination of block set "GTCC" shows that "GTCC" occurs upstream of "CTTG" and "TACC". However, an examination of block set "CTTG" and an examination of block set "TACC" indicates that sequence blocks "CTTG" and "TACC" can both occur upstream and downstream of each other, which appears to conflict with the observation that these sequence blocks only occur once in the sequence of strand B. There is actually no conflict. Each of these sequence blocks does indeed occur only once. It is just that their positions, relative to one another, in strand B are obscured by the presence of conflicting information from the relative positions of oligos that occur in strand A. This ambiguity (indicated by the identical positions of sequence blocks "CTTG" and "TACC" in the diagram to the left of the arrow in Figure 11e) is resolved by the remainder of the information. The positions of those sequence blocks that can potentially occur more than once in the sequence of strand B is determined from other block sets. First, the block sets of the sequence blocks that definitely occur only once in the sequence (namely, block sets "CTTG", "GTCC", and "TACC") are consulted. The range of positions at which these other sequence blocks can occur (relative to the positions of other blocks) is indicated in the diagram to the left side of the arrow in Figure 11e.

The assembly of the nucleotide sequence of Strand B proceeds as follows: "ATG" is upstream of all other blocks. The uniquely occurring block immediately downstream of "ATG" is "GTCC". "ATG" and "GTCC" cannot be directly joined. However, "ATG" can be directly joined to "TGGT", so the correct order is to join "ATG" to "TGGC", and then to join "TGGC" to "GTCC". Neither "CTTG" nor "TACC" can be directly joined to "GTCC". Three different sequence blocks can be used to bridge this gap (namely, "CCT", "GTA", and "TGGT"). The only combination of these three sequence blocks that can fill this gap is "CCT" alone, which bridges the gap between "GTCC" and "CTTG". This resolves the ambiguity as to the relative positions of "CTTG" and "TACC". "CTTG" is therefore upstream of "TACC". "CTTG" cannot be directly joined to "TACC". Again, there are three different sequence blocks that can be used

to fill this gap (namely, "CCT", "GTA", and "TGGT"). The only combination of these three sequence blocks that can fill this gap is "TGGT" and "GTA" (i.e.,"GTTG" is joined to "TGGT", "TGGT" is joined to "GTA", and "GTA" is joined to "TACC"). And finally, "CTA", which occurs upstream of all other blocks, must be included in the sequence. However, "TACC" cannot be directly joined to "CTA". There are three different sequence blocks that can be used to fill this gap (namely, "CCT", "GTA", and "TGGT"). The only combination of these three sequence blocks that can fill this gap is "CCT" alone. Thus, the assembly of the sequence of Strand B from its sequence blocks is completed. Note that some sequence blocks that could potentially occur in the sequence more than once, actually occur only once (e.g., "GTA"), while others actually occur more than once (e.g., "CCT").

Using the methods of this invention, the entire sequence of strand B is unambiguously determined, despite the fact that some oligos occur more than once in its sequence, despite the fact that more than one sequence block can be assembled from the oligos that occur in the strand, despite the fact that the multiplicity of occurrence of each oligo is not determined during surveying, despite the fact that the strand is analyzed in a mixture of strands, and despite the fact that the other strand in the mixture possesses many of the same oligos.

5. Uses of sectioned oligonucleotide arrays for manipulating nucleic acids --

In the examples described below, it is assumed that the sequences of the nucleic acids to be manipulated have already been established. It is not necessary, in these manipulations, that the sample be distributed across the entire array. Instead, a sample can be delivered directly to the well in the array where a particular oligo (or a particular strand) is immobilized. The arrays enable a large number of specifically directed manipulations of nucleic acids to be carried out.

5.1. Cleavable primers --

Amplification of strands and partials following separation (or generation) on a sectioned array requires that their ends be provided with priming regions. The priming regions can be undesirable in subsequent use, such as the making of recombinants or site-directed mutants. For some uses it is desirable to substitute new priming regions for the old. For those uses, the primers used for amplification must first be removed from the 5' ends.

Where the junction of the primer and the strand is contained within a unique restriction site, the primer can be removed by treating a double-stranded version of the strand with a corresponding restriction endonuclease. However, restriction sites will often not be present at the junctions. A solution to this problem is to make the primer (or even only the junction nucleotide in the primer) chemically different from the rest of the strand. The primer in these examples resides at the strand's 5' terminus.

5.1.1. Cleavage of primers by alkaline hydrolysis or by ribonuclease digestion --

This method is suitable for removal of oligoribonucleotide primers, or mixed RNA/DNA primers whose 3' terminal nucleotide (which becomes a junction nucleotide upon primer extension) is a ribonucleotide. Such primers are incorporated at the 5' end of DNA strands or partials during amplification.

Alkaline hydrolysis cleaves a phosphodiester bond that is on the 3' side of a ribonucleotide, and leaves intact a phosphodiester bond that is on the 3' side of a deoxyribonucleotide. After alkaline hydrolysis, the pH of the reaction mixture is returned to a neutral value by the addition of acid, and the sample can be used without purification. Primers containing a riboadenylate or a riboguanylate residue at their 3' end can effectively be removed from a DNA strand or partial by treatment with T_2 ribonuclease. After treatment, the sample is heated to 100°C to inactivate the ribonuclease, and can be used without purification. In both these cases, the released 5' terminus of

the strand (or partial) is left dephosphorylated. Therefore, if the strand obtained is subsequently used for ligation, it should be phosphorylated by incubation with polynucleotide kinase.

5.1.2. Cleavage of primers from DNA strands (or partials) synthesized from phosphorothioate nucleotide precursors --

In this method, oligodeoxynucleotide or oligoribonucleotide primers are synthesized from natural nucleotides, but strand amplification is carried out in the presence of only α-phosphorothioate nucleotide precursors. Subsequent digestion of the synthesized strands with a 5'-3' exonuclease, such as calf spleen 5'-3' exonuclease, results in the elimination of all primer nucleotides except the original 3'-terminal (junction) nucleotide of the primer, with the released 5'-terminal group of a strand or partial being unphosphorylated. The junction nucleotide is not removed, because it is joined to the rest of the strand by a phosphorothioate diester bond. Therefore, the strand obtained has an extra nucleotide at its 5' end. This does not present a problem when the presence of the former junction nucleotide at the 5' end of the strand is compatible with the subsequent use of the strand. The presence of the extra nucleotide can also be useful for site-directed mutagenesis.

If the primer-deprived strand so obtained is to be ligated, the use of spleen exonuclease, which leaves 5'-hydroxyl groups, must be then followed by phosphorylation with polynucleotide kinase. Therefore, where the strand is to be ligated, the use of bacteriophage lambda or bacteriophage T7 5'-3' exonuclease is preferable over spleen exonuclease, since they leave 5'-phosphoryl groups at the site of cleavage.

5.2. Generation of recombinant nucleic acids --

In the method described below, two nucleic acid strands are ligated in one round of ligation. It is possible to keep repeating the process any desired number of times to ligate the desired number of strands.

In this example, a sectioned array contains immobilized oligos that consist of two portions, one complementary to the 3'-

terminal sequence of one of the moieties to be ligated, and the other complementary to the 5'-terminal sequence of the other moiety to be ligated. The immobilized oligos can have either free 3' or 5' ends. The relevant termini of the moieties to be ligated should be deprived of priming regions, but priming regions (preferably different) should be preserved at the opposite termini to allow amplification of the recombinants. After hybridization in an appropriate well, the two nucleic acid strands are ligated to each other utilizing DNA ligase. Unligated strands are then washed away. Only ligated strands possess two terminal priming regions required for PCR. The strands that are to be ligated can be used in a mixture with other strands, provided that no other strands have with the same oligos at the termini deprived of priming regions.

Many different strands can be ligated to one particular strand (or partial), to produce many recombinant variations of one gene. In that case, one portion of the splint, i.e., the immobilized oligo is a constant segment, and the other portion is a variable segment, i.e., a binary array is used. The constant segment binds to the strand to be included in every recombinant, and the variable segment binds to the end of a strand to be fused with the invariant strand.

5.3. Site-directed mutagenesis --

The ability to prepare any partial of a strand according to the invention provides the opportunity to make nucleotide substitutions, deletions and insertions at any chosen position within a nucleic acid. Moreover, the use of sectioned arrays makes it possible to perform site-directed mutagenesis at a number of positions (even at all positions) at once, and in a particular embodiment, to determine, within individual wells of the array, properties of the encoded mutant proteins.

Mutations are introduced into a strand by first preparing partials having variable ends that correspond to the segment to be mutated, that segment preceding the location of the intended mutation. Then mutagenic nucleotides or oligos are introduced into the variable ends. The mutated partials are then extended

the length of the full sized strand using the complementary copy of the original non-mutated strand as a template.

In this method, complements of partials (i.e., strands whose 5' termini are variable and 3' termini are fixed) are used. Their 5'-terminal priming regions are removed and then phosphorylated by incubation with polynucleotide kinase, and the partials are then ligated by incubation with RNA ligase to the free 3' hydroxyls of oligoribonucleotides immobilized on a 3' sectioned ordinary array. The sequence of the immobilized oligo to which a partial is ligated is identical to the oligo segment that occurs in the original (full-length) strand immediately adjacent to the end of the partial, except for one (or a few) nucleotide difference(s) that corresponds to mutation(s) to be introduced.

The nucleotide differences are preferably located at the 3' terminus of the immobilized oligo, and can correspond to a nucleotide substitution, insertion, or deletion. A deletion can be of any size. For a large insertion, the ligated partial, or the immobilized oligo, can first be fused to a nucleic acid containing all or part of the sequence to be inserted.

After washing away material not covalently bound, the immobilized strand is linearly copied, taking advantage of the priming region at its (fixed) 3' end. The copies correspond to partials that have been extended by the oligos containing the mutation(s). The copies are annealed to their complementary full-length strands, and their 3' termini extended by incubation with DNA polymerase, using the parental strand as a template. Finally, the extended mutant strands are amplified by PCR. It is important that the primers utilized for amplification of a partial used for mutagenesis be different from the primers used to amplify the original (non-mutant) full-length strand. This assures that only mutant strands are amplified.

We claim:

1. A binary oligonucleotide array comprising an array of predetermined areas on a surface of a solid support, each area having therein, covalently linked to said surface, multiple copies of a binary oligonucleotide of a predetermined sequence, said binary oligonucleotide consisting of a constant nucleotide sequence adjacent to a variable nucleotide sequence, wherein the constant nucleotide sequence is the same for all oligonucleotides in the array.

2. A binary array according to claim 1 wherein the binary oligonucleotides consist of deoxyribonucleotides.

3. A binary array according to claim 1 wherein the binary oligonucleotides consist of ribonucleotides.

4. A binary array according to claim 1 wherein one or more of nucleotides of the binary oligonucleotides are modified.

5. A binary array according to claim 1 wherein one or more of the nucleotides of the binary oligonucleotides are non-standard.

6. A binary array according to claim 1 wherein the binary oligonucleotides are mixed.

7. A comprehensive binary array according to claim 1

8. A comprehensive binary array according to claim 7 wherein the binary oligonucleotides in each area have variable sequences of the same length.

9. A 3' binary array according to claim 1.

10. A 5' binary array according to claim 1.

11. A 3' binary array according to claim 9, wherein each covalently linked binary oligonucleotide has its constant sequence adjacent to the 5' end of its variable sequence.

12. A 5' binary array according to claim 10, wherein each covalently linked binary oligonucleotide has its constant sequence adjacent to the 3' end of its variable sequence.

13. A binary array according to claim 2 wherein all or part of the constant nucleotide sequence is complementary to a predetermined restriction recognition sequence.

14. A binary array according to claim 1 having an oligonucleotide hybridized to all or part of the constant sequence which is ligatable to the terminus of an adjacent nucleic acid hybridized to the oligonucleotide.

15. In an oligonucleotide array having variable-sequence oligonucleotides immobilized in a predetermined pattern of areas on a solid support, the improvement comprising including in said oligonucleotides a constant sequence of predetermined length.

16. A sectioned binary array according to claim 1.

17. A comprehensive sectioned binary array according to claim 16.

18. A 3' binary oligonucleotide array according to claim 17, wherein each covalently linked binary oligonucleotide has its variable sequence adjacent to the 5' end of its constant sequence.

19. A 5' binary oligonucleotide array according to claim 17, wherein each covalently linked binary oligonucleotide has its variable sequence adjacent to the 3' end of its constant sequence.

20. A binary oligonucleotide array according to claim 1, wherein said constant nucleotide sequence comprises one or more functional sequences selected from the group consisting of a nucleic acid polymerase priming region, an RNA polymerase promoter region, and a restriction endonuclease recognition site.

21. A binary oligonucleotide array according to claim 20, wherein said functional sequence is a priming region.

22. A binary oligonucleotide array according to claim 1, wherein each binary oligonucleotide is covalently linked to said surface through a long polymer chain.

23. A binary oligonucleotide according to claim 2, wherein said deoxyribonucleotides comprise at least one modified nucleotide.

24. A sectioned oligonucleotide array comprising an array of predetermined areas on a surface of a solid support, each area having therein, covalently linked to said surface multiple copies of an oligonucleotide, wherein said areas are physically separated from one another into sections, such that nucleic acids in an aqueous solution generated in one section cannot migrate to another section.

25. A sectioned oligonucleotide array according to claim 24 further comprising a lattice attached to said surface.

26. A sectioned oligonucleotide array according to claim 25, wherein said lattice is removably attached to said surface.

27. A sectioned oligonucleotide array according to claim 25, further comprising a cover removably attachable to said lattice.

28. A sectioned oligonucleotide array according to claim 24, wherein said sections comprise wells in said solid support.

29. A sectioned oligonucleotide array according to claim 28, further comprising a cover removably attachable to said solid support.

30. A sectioned oligonucleotide array according to claim 24, comprising a gel which physically separates said areas by preventing nucleic acids in an aqueous solution placed in one area from migrating to another area.

31. A sectioned oligonucleotide array according to claim 24, wherein said sections are mechanically separated from one another.

32. A sectioned oligonucleotide array according to claim 27, wherein said cover comprises a replica array.

33. A sectioned oligonucleotide array according to claim 29, wherein said cover comprises a replica array.

34. A sectioned array according to claim 24 wherein all of the oligonucleotides in individual areas are of the same sequence.

35. A sectioned array according to claim 24 wherein not all oligonucleotides in each area are of the same sequence.

36. A method of sorting a mixture of nucleic acid strands comprising the steps of:

a) providing a solution containing a mixture of nucleic acid strands in single-stranded form and

b) contacting said solution to a first binary oligonucleotide array of predetermined areas on a surface of a solid support, each area having therein, covalently linked to said surface, copies of a binary oligonucleotide, said binary oligonucleotide consisting of a constant nucleotide sequence adjacent to a variable nucleotide sequence, wherein the constant nucleotide sequence is the same for all oligonucleotides in the array, wherein said step of contacting is carried out under conditions

promoting perfect hybridization of said strands to said binary oligonucleotides.

37. A method according to claim 36 wherein said array is comprehensive.

38. A method according to claim 36 wherein said array is a 3' array.

39. A method according to claim 36 wherein said binary oligonucleotides are complementary to sequences that possibly occur in the strands in said mixture.

40. A method according to claim 39 wherein said array is comprehensive.

41. A method according to claim 36 wherein said array is a sectioned array, further comprising the step of amplifying strands hybridized in at least some of said areas to produce copies of said hybridized strands.

42. A method according to claim 36 further comprising removing strands that have not perfectly hybridized.

43. A method according to claim 42 further comprising adding a terminal extension to at least one terminus of the strands, said terminal extension having a sequence which substantially does not occur in the strands.

44. A method according to claim 43 wherein a terminal extension is added to the strands by ligation of hybridized strands to masking oligonucleotides, said masking oligonucleotides being also hybridized to said binary oligonucleotides.

45. A method according to claim 44 wherein a second terminal extension is added to the strands prior to said step of contacting, said second terminal extension being added to termini

not hybridized to said binary oligonucleotides during said step of contacting.

46. A method according to claim 42 further comprising releasing hybridized strands on a sectioned array into solution without mixing of material in said areas and rebinding them to said binary oligonucleotides followed by removing unhybridized strands.

47. A method according to claim 42 further comprising releasing hybridized strands in solution and rebinding to a replica array followed by removing unhybridized strands.

48. A method according to claim 42 wherein the mixture of nucleic acid strands comprises RNA.

49. A method according to claim 42 wherein the mixture of nucleic acid strands is comprised of DNA fragments obtained by site specific degradation.

50. A method according to claim 43 wherein the mixture is comprised of DNA fragments obtained by digestion with a restriction endonuclease and wherein the constant region of the binary oligonucleotide contains the complement of the restriction endonuclease recognition site, and wherein addition of the terminal extension restores the recognition site.

51. A method according to claim 42 further comprising generating complementary copies of hybridized strands.

52. A method according to claim 51 wherein the array is a 3' array wherein each binary oligonucleotide has its variable sequence adjacent to the 5' end of its constant sequence, and the copies are generated using a DNA polymerase and using the binary oligonucleotide as a primer.

53. A method according to claim 51 wherein the array is a 5' array wherein each binary oligonucleotide has its variable sequence adjacent to the 3' end of its constant sequence, and the copies are generated using a DNA polymerase using a primer hybridized to a 3' terminal extension of the hybridized strands, and the copies are then ligated to the 5' end of the binary oligonucleotides.

54. A method according to claim 44 further comprising amplifying the hybridized strands.

55. A method according to claim 51 further comprising removing the hybridized strands and amplifying the complementary copies of the hybridized strands.

56. A method according to claim 55 wherein the hybridized strands have 3' and 5' terminal extensions, and the amplification is a polymerase chain reaction.

57. A method according to claim 55 wherein the hybridized strands have a terminal extension and the amplification is linear.

58. A method according to claim 36 wherein said step of providing comprises digesting genomic DNA with a restriction endonuclease to create DNA fragments;

(a) modifying said fragments by adding a first constant sequence to their strands' 3' termini and a second constant sequence to their strands' 5' termini to create priming regions including restored restriction sites; and

(b) denaturing the modified fragments to form a mixture of single nucleic acid strands.

59. A method according to claim 58 wherein said array is a sectioned, comprehensive array, further comprising the step of amplifying strands hybridized in said areas by symmetric PCR.

60. A method according to claim 58 further comprising the step of amplifying said mixture of single nucleic acid strands by asymmetric PCR.

61. A method according to claim 36 wherein said binary oligonucleotides or portions thereof are complementary to terminal sequences that possibly occur in one end of the strands in said mixture and that are substantially non-complementary to internal sequences in the strands in said mixture.

62. A method according to claim 61 wherein said array is a sectioned array, further comprising the step of amplifying strands hybridized in at least some of said areas to produce amplified copies of said single nucleic acid strands.

63. A method according to claim 62 wherein said array is a comprehensive array.

64. A method according to claim 62 wherein said array is a 3' array.

65. A method according to claim 61 wherein said step of providing comprises digesting genomic DNA with a restriction endonuclease to create DNA fragments, modifying said fragments by adding a first constant sequence to their strands' 3' termini to create priming regions including restored restriction sites, and denaturing the modified fragments into a mixture of single nucleic acid strands.

66. A method according to claim 61 wherein said step of providing comprises digesting genomic DNA with a restriction endonuclease to create DNA fragments;

(a) modifying said fragments by adding a first constant segment to one of their strands' 3' and 5' termini to create priming regions including restored restriction sites; and

(b) denaturing the modified fragments into a mixture of denatured nucleic acid strands each having a priming region only at one end.

67. A method according to claim 66 wherein said first binary sorting array is a 3' array.

68. A method according to claim 67 further comprising the steps of

(a) generating an immobilized copy of each strand hybridized to the array by incubation with a DNA polymerase using the immobilized oligonucleotide as a primer and a hybridized strand as a template; and

(b) washing to remove from the array all materials not covalently bound to the array.

69. A method according to claim 68, wherein said step of modifying comprises adding a first constant sequence to their strands' 5' termini and wherein said 3' array contains binary oligonucleotides to which are hybridized masking oligonucleotides, further comprising the steps of

(a) ligating said masking oligonucleotides to denatured nucleic acid strands hybridized to said binary oligonucleotides such that their 3' termini are immediately adjacent to one of said masking oligonucleotides, and

(b) washing under conditions such that only strands so ligated will remain.

70. A method according to claim 69 wherein said step of adding a first constant sequence includes ligation of a double-stranded oligodeoxyribonucleotide adaptor.

71. A method according to claim 69 wherein said step of adding a first constant sequence includes ligation of a single-stranded oligoribonucleotide.

72. A method according to claim 68 wherein said step of modifying comprises adding a first constant sequence to their strands' 3' termini.

73. A method according to claim 72 wherein said first constant sequence is a homopolynucleotide tail added by extension of the strands' 3' termini by enzymatic extension.

74. A method according to claim 72 further comprising the step of adding a second constant sequence to the 3' termini of the immobilized copies.

75. A method according to claim 74 wherein said second constant sequence is a homopolynucleotide tail added by extension of said immobilized copies' 3' termini by enzymatic extension.

76. A method according to claim 68 wherein said first binary oligonucleotide array is a sectioned array, further comprising the step of amplifying said washed, immobilized copies to produce amplified copies.

77. A method according to claim 76 wherein said step of amplifying comprises PCR.

78. A method according to claim 76 wherein said first binary oligonucleotide array is a comprehensive array.

79. A method according to claim 76 further comprising contacting said amplified copies from at least one area of said 3' array to a second binary oligonucleotide array containing immobilized binary oligonucleotides whose constant sequence is identical or complementary to the 3' terminus of the immobilized copies.

80. A method according to claim 62 further comprising contacting said amplified copies from at least one area of said first binary oligonucleotide array to a second binary oligonucleotide array containing immobilized binary oligonucleotides that are com-

plementary to terminal sequences that possibly occur in either the other ends of said denatured nucleic acid strands or the complements of said other ends, and that are not complementary to internal sequences in the strands in said mixture or their complements.

81. A method according to claim 61 wherein said step of providing comprises digesting genomic DNA with a restriction endonuclease to create DNA fragments, and denaturing said fragments into a mixture of denatured nucleic acid strands.

82. A method according to claim 81 wherein said first binary oligonucleotide array is a 3' array containing binary oligonucleotides to which are hybridized masking oligonucleotides, further comprising the steps of ligating said masking oligonucleotides to denatured nucleic acid strands hybridized to said binary oligonucleotides such that their 3' termini are immediately adjacent to one of said masking oligonucleotides, washing under conditions such that only strands so ligated will remain, and generating an immobilized copy of each ligated strand by incubation with a DNA polymerase.

83. A method according to claim 82 further comprising the steps of adding a constant sequence to the 5' termini of the hybridized strands by ligation of a single-stranded oligoribonucleotide; incubating with a DNA polymerase to extend the immobilized copies; washing to remove from the array all materials not covalently bound to the array; and amplifying said washed, immobilized copies to produce amplified copies.

84. A method according to claim 83 wherein said step of amplifying comprises PCR.

85. A method according to claim 83 wherein said first sorting array is a comprehensive array.

86. A method according to claim 83 further comprising contacting said amplified copies from at least one area of said 3' array to a second binary array containing immobilized binary oligonucleotides whose constant sequence is identical or complementary to the 3' terminus of said immobilized copies.

87. A method according to claim 67 further comprising the steps of adding a constant sequence to the 3' termini of the immobilized copies by enzymatic extension thereof; washing to remove from the array all materials not covalently bound to the array; and amplifying said washed, immobilized copies to produce amplified copies.

88. A method according to claim 87 wherein said step of amplifying comprises PCR.

89. A method according to claim 87 wherein said first sorting array is a comprehensive array.

90. A method according to claim 87 further comprising contacting said amplified copies from at least one area of said 3' array to a second terminal binary array containing immobilized binary oligonucleotides whose constant sequence is identical or complementary to the 3' terminus of said immobilized copies.

91. A method according to claim 61 wherein said step of providing comprises digesting genomic DNA with a site-specific cleaving agent to create DNA fragments.

92. A method according to claim 91 wherein said agent is an endonuclease.

93. A method according to claim 91 wherein said agent is a chemical agent.

94. A method according to claim 61 wherein said nucleic acid strands are cDNA strands.

95. A method according to claim 61 wherein said nucleic acid strands are RNA strands.

96. A method according to claim 95 wherein said RNA strands are eukaryotic mRNA strands, and wherein said step of providing comprises removing 5'-cap structures.

97. A method according to claim 95 wherein said RNA strands lack a poly(A) tail.

98. A method according to claim 61 wherein said step of providing comprises digesting genomic DNA with a restriction endonuclease to create DNA fragments;

(a) modifying said fragments by adding a first constant sequence to their strands' 3' termini and a second constant sequence to their strands' 5' termini to create priming regions including restored restriction sites; and

(b) denaturing the modified fragments into a mixture of single nucleic acid strands.

99. A method according to claim 98 wherein the 3' priming regions are complementary to the 5' priming regions.

100. A method according to claim 99 wherein said array is a 3' array, further comprising the steps of

(a) generating an immobilized copy of each strand hybridized to the array by incubation with a DNA polymerase; and

(b) washing to remove from the array all materials not covalently bound to the array.

101. A method according to claim 100 wherein said array is a sectioned array, further comprising the step of amplifying strands hybridized in at least some areas by PCR to produce amplified copies of each said immobilized copy.

102. A method according to claim 101 wherein said array is a comprehensive array.

103. A method according to claim 99 wherein addition of said first constant sequence and said second constant sequence includes ligation of a double-stranded oligodeoxyribonucleotide adaptor to the strands' 5' termini.

104. A method according to claim 99 wherein addition of said first constant sequence and said second constant sequence includes ligation of a single-stranded oligonucleotide to the strands' 5' termini.

105. A method according to claim 99 wherein addition of said first constant sequence and said second constant sequence includes enzymatic extension of the strands' 3' termini by the synthesis of a homopolynucleotide tail.

106. A method according to claim 101 further comprising contacting said amplified copies from at least one areas of said 3' array to a second binary array under conditions promoting hybridization of said amplified copies to the binary oligonucleotides in said second array.

107. A method according to claim 106 wherein said amplified copies are produced by symmetric PCR and wherein said second array is a 3' array.

108. A method according to claim 106 wherein said first array and said second array are comprehensive.

109. The product of a method according to claim 100.

110. A method of sorting a mixture of nucleic acid strands comprising the steps of

a) providing a solution containing a mixture of nucleic acid strands in single stranded form, and

b) contacting said solution to an oligonucleotide array of predetermined areas on a surface of a solid support, each area having therein copies of an immobilized oligonucleotide, the nucleotide sequence of immobilized oligonucleotides in separate areas being different, wherein said contacting is performed under conditions that promote the formation of perfect hybrids.

111. A method according to claim 110 wherein said array is comprehensive.

112. A method according to claim 110 wherein the array is sectioned.

113. A method according to claim 110 wherein the immobilized oligonucleotides are between 6 and 30 nucleotides long.

114. A method according to claim 110 wherein the array is a 3' array.

115. A method according to claim 110 wherein the array is a 5' array.

116. In a method wherein two nucleic acid strands are ligated to each other in order to form a recombinant product, the improvement comprising hybridizing first nucleic acid strands to immobilized oligonucleotides in an oligonucleotide array prior to ligation to second nucleic acid strands, said oligonucleotide array comprising an array of predetermined areas on a surface of a solid support, each area having copies of an oligonucleotide immobilized thereon.

117. A method according to claim 116 wherein the first nucleic acid strands have different nucleotide sequences in each area of the array.

118. A method according to claim 116 wherein the second nucleic acid strands have different nucleotide sequences in each area of the array.

119. A method according to claim 116 wherein the array is a comprehensive array.

120. A method according to claim 116 wherein the oligonucleotides immobilized in each area are of the same length.

121. A method according to claim 116 wherein the oligonucleotides consist of the group consisting of deoxyribonucleotides, ribonucleotides, mixed deoxyribonucleotides and ribonucleotides, modified deoxyribonucleotides, modified ribonucleotides, and non-standard nucleotides.

122. A method according to claim 116 wherein the second nucleic acid strands are not also hybridized to the immobilized oligonucleotides.

123. A method according to claim 122 wherein the second nucleic acid strands are strands of double stranded nucleic acids.

124. A method according to claim 123 wherein the set of double stranded nucleic acids has one end adapted for ligation to blunt ends formed by hybridization of the first nucleic acids to the immobilized oligonucleotides.

125. A method according to claim 116 wherein non-ligating termini of the first nucleic acid strands and the double stranded nucleic acids contain priming regions for amplification.

126. A method according to claim 125 wherein following ligation of the first nucleic acids to the second nucleic acids, polymerase chain reaction amplification is carried out.

127. A method according to claim 124 wherein the double stranded nucleic acids are ligated to the immobilized oligonucleotide using RNA ligase prior to ligation of the first nucleic acid strands and the second nucleic acid strands.

128. A method according to claim 123 wherein the second set of nucleic acids is the same in every area of array.

129. A method according to claim 123 wherein the first nucleic strands are hybridized to the immobilized oligonucleotides while contained in a mixture of one or more different strands, said different strands having terminal sequences different from corresponding termini to be ligated of the first nucleic acid strands.

130. A method according to claim 116 wherein both the first nucleic acid strands and the second nucleic acid strands are hybridized to the immobilized oligonucleotides in the array prior to ligation.

131. A method according to claim 130 wherein both the first and second nucleic acid strands contain priming regions at their non-ligating termini.

132. A method according to claim 131 wherein the first and second nucleic acid strands are amplified in a polymerase chain reaction following ligation.

133. A method according to claim 130 wherein both the first and second nucleic acids are, prior to hybridization to the immobilized oligonucleotides, contained in mixtures of nucleic acids having terminal sequences different from the corresponding termini to be ligated of the first nucleic acid strands and the second nucleic acid strands.

134. A method according to claim 36 further comprising sorting the hybridized nucleic acid strands or their copies in an area of

the first binary array by contacting them to a second oligonucleotide array.

135. A method according to claim 134 wherein the strands or their copies are contacted to all areas of the array.

136. A method according to claim 36 wherein the nucleic acid strands are contacted to all areas of a second binary array.

137. A method according to claim 134 wherein cleavable primers are used following said step of contacting for amplification of hybridized strands.

138. A method according to claim 137 further comprising cleaving the cleavable primers from the strands and adding new terminal extensions.

139. A method according to claim 134 wherein the contents of an area of the first binary array are contacted with only predetermined areas of a second binary array.

140. A method according to claim 36 further wherein contents in an area of the binary array are contacted with the corresponding area of a replica array.

141. A method according to claim 134 wherein the second oligonucleotide array is a second binary array.

142. A method for introducing a site directed mutation into a nucleic acid strand on an oligonucleotide array using a partial, said partial corresponding to a region of the nucleic acid strand adjacent to the location of the site directed mutation to be introduced, comprising the steps:

(a) separately ligating said partial to the free terminus of a preselected immobilized oligonucleotide in the oligonucleotide array to obtain a mutated partial, said oligonucleotide array comprising an array of predetermined areas on the

surface of a solid support, each area having therein a preselected immobilized oligonucleotide, said preselected oligonucleotide having a sequence adapted to introduce a mutation to the partial added to the area; and

(b) generating, using the mutated partial, a nucleic acid containing the mutation.

143. A method according to claim 142 wherein step b is accomplished by

(a) hybridizing a complementary copy of the mutated partial to a template having the complementary sequence of the terminal portion of the nucleic acid strand which is not contained in the partial; and

(b) carrying out a polymerase reaction, a ligation reaction or both a polymerase reaction and ligation reaction to join the remaining region of the nucleic acid strand to the mutated partial.

144. A method for making immobilized partial copies of a nucleic acid strand on a 3' or 5' oligonucleotide array, comprising the steps:

(a) hybridizing the strand to the array by an oligonucleotide segment contained in the strand, said array comprising predetermined areas on a surface of a solid support, each area having therein immobilized oligonucleotides consisting of a predetermined variable sequence, said hybridization taking place under conditions that promote the formation of perfect hybrids of the length of the immobilized oligonucleotide in each area, and

(b) where the strand is hybridized to a 3' array, enzymatically extending the immobilized oligonucleotide using the hybridized strand as a template, and where the strand is hybridized to a 5' array, hybridizing a primer to a priming region contained in the 3' terminus of the hybridized strand, then enzymatically extending the primer to form an extension product, then ligating the extension product to the immobilized oligonucleotide.

145. A method according to claim 144 wherein the strand is hybridized to a 3' array, further comprising amplifying the immobilized partial copies using a primer or promoter complement appropriate to hybridize to a priming region or promoter sequence at the immobilized partial copies' 3' termini, and an appropriate polymerase.

146. A method according to claim 144 wherein the oligonucleotide array is substantially comprehensive.

147. A method according to claim 146 wherein a substantially complete set of immobilized partial copies is generated on the array by

(a) hybridizing the strand to the array by substantially all oligonucleotides present in the strand;

(b) performing step (b) on all hybridized strands.

148. A method according to claim 146 wherein a substantially complete set of amplified partials is generated on a 3' array by

(a) hybridizing the strand to the 3' array by substantially all oligonucleotides present in the strand;

(b) performing step (b) on all hybridized strands; and

(c) amplifying substantially all immobilized partial copies by using a primer or promoter complement appropriate to hybridize to a priming region or promoter sequence at the partial copy's fixed terminus, and an appropriate polymerase.

149. A method according to claim 148 wherein following step (a) unhybridized and imperfectly hybridized strand copies are removed.

150. A method according to claim 149 wherein the array is sectioned.

151. A method according to claim 150 wherein the strand is contained in a mixture of strands which are subjected to the same steps on the array.

152. A method according to claim 151 wherein the priming region is a terminal extension introduced in all strands in the mixture.

153. A method according to claim 149 wherein the priming region or promoter is added to the 5' terminus of the nucleic acid strand prior to hybridizing the strand to the array.

154. A method according to claim 150 further wherein the oligonucleotide content in an area of the array is surveyed.

155. The product of a method according to claim 144.

156. The product of a method according to claim 146.

157. A method according to claim 144 wherein the strand is contained in a mixture of sorted strands subjected to the method, said mixture of sorted strands being from an area of a sorting array.

158. A method according to claim 157 further wherein mixtures of strands from different areas of the sorting oligonucleotide array are hybridized to the 3' or 5' oligonucleotide array.

159. A method according to claim 144 wherein the nucleic acid is a previously prepared partial.

160. A method according to claim 145 further comprising sorting partials or their copies from an area of the oligonucleotide array on a second oligonucleotide array.

161. A method according to claim 145 further comprising sorting partials or their copies from an area of the oligonucleotide array according to variable sequences adjacent their fixed ends on a binary oligonucleotide array.

162. A method of claim 144 further comprising ligating a partial or its copy in single stranded or double stranded form to a second nucleic acid strand.

163. A method according to claim 162 wherein the second nucleic acid strand is a previously obtained partial.

164. A method according to claim 145 further wherein a cleavable primer, at an end of a partial to be ligated, is used for amplification, and further comprising cleaving the primer and then ligating the partial to a second nucleic acid strand.

165. A method according to claim 162 further comprising exponentially amplifying ligated product using priming regions at non-ligated termini.

166. A method according to claim 165 further wherein the priming regions at the non-ligated termini of the ligated product are adapted to permit amplification only of the ligated product.

167. A method according to claim 144 further wherein a partial obtained is ligated to an oligonucleotide or to a second nucleic acid strand adapted to introduce a site directed mutation, with respect to the nucleic acid strand that the partial was generated from, at the ligated terminus of the partial.

168. A method according to claim 167 wherein the oligonucleotide is immobilized in a second oligonucleotide array.

169. A method for sorting partials by their variable termini on a binary oligonucleotide array, which partials have been prepared by random chemical or enzymatic degradation of one or more nucleic acid strands, said binary array comprising an array of predetermined areas on a surface of a solid support, each area having therein copies of a binary oligonucleotide of a predetermined sequence, said binary oligonucleotide consisting of a constant nucleotide sequence adjacent to a variable nucleotide

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/02 — A1

(11) International Publication Number: **WO 99/07888**

(43) International Publication Date: 18 February 1999 (18.02.99)

(21) International Application Number: PCT/US98/15408

(22) International Filing Date: 24 July 1998 (24.07.98)

(30) Priority Data:
08/906,543 5 August 1997 (05.08.97) US

(71) Applicant *(for all designated States except US)*: PRESIDENT & FELLOWS OF HARVARD COLLEGE [US/US]; University Place, 124 Mt. Auburn Street, Cambridge, MA 02138 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: CHURCH, George, M. [US/US]; 218 Kent Street, Brookline, MA 02146 (US). BULYK, Martha, L. [US/US]; 103 Browne Street #3, Brookline, MA 02146 (US).

(74) Agent: WILLIAMS, Kathleen, M.; Banner & Witcoff, Ltd., 28th floor, 28 State Street, Boston, MA 02111 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: SURFACE-BOUND, BIMOLECULAR, DOUBLE-STRANDED DNA ARRAYS

(57) Abstract

The invention provides an array of surface-bound, bimolecular, double-stranded, nucleic acid molecules, the array comprising a solid support, and a plurality of different double-stranded nucleic acid molecule members, a member comprising a first nucleic acid strand linked to the solid support and a second nucleic acid strand which is substantially complementary to the first strand and complexed to the first strand by Watson-Crick base pairing, wherein at least a portion of the members have a second nucleic acid strand is substantially complementary to and base paired with the first strand along the entire length of the first strand.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

SURFACE-BOUND, BIMOLECULAR, DOUBLE-STRANDED DNA ARRAYS

FIELD OF INVENTION

The invention relates to nucleic acid arrays.

BACKGROUND OF THE INVENTION

Compact arrays or libraries of surface-bound, double-stranded oligonucleotides are of use in rapid, high-throughput screening of compounds to identify those that bind, or otherwise interact with, short, double-stranded DNA sequence motifs. Of particular interest are proteins, particularly *trans*-regulatory factors, that control gene transcription. Ideally, such an oligonucleotide array is bound to the surface of a solid support matrix that is of a size that enables laboratory manipulations, e.g. an incubation of a candidate protein with the nucleic acid targets sequences thereon, and that is itself inert to chemical interactions with experimental proteins, buffers and/or other components. In addition, it is desirable that the absolute number of unique target sequences in the array be maximized, since methods of high-throughput screening are used in the attempt to minimize repetition of steps that are labor-intensive or otherwise costly.

A high-density, double-stranded DNA array complexed to a solid matrix is described by Lockhart (U.S. Patent No.: 5,556,752); however, the DNA molecules therein disclosed are produced as unimolecular products of chemical synthesis. Each member of the array contains regions of self-complementarity separated by a spacer (*i.e.* a single-strand loop), such that these regions hybridize to each other in order to produce a double-helical region. A difficulty of such a production method arises when the accuracy of chemical synthesis is considered in light to that of that demonstrated by proteinaceous DNA polymerase molecules. It is estimated that enzymatic synthesis of second-strand DNA from a first-strand template operates at 100-fold higher fidelity than do chemical synthetic procedures. Further, it is required that those regions of complementary nucleic acid sequences that must hybridize in order to form the double-helical structure are physically attached to each other by a linker subunit.

SUMMARY OF THE INVENTION

The present invention encompasses an array of surface-bound, bimolecular, double-stranded, nucleic acid molecules, the array comprising a solid support, and a plurality of different double-stranded nucleic acid molecule members, a member comprising a first

nucleic acid strand linked to the solid support and a second nucleic acid strand which is substantially complementary to the first strand and complexed to the first strand by Watson-Crick base pairing, wherein at least a portion of the members have a second nucleic acid strand which is substantially complementary to and base paired with the first strand along the entire length of the first strand.

The term "synthetic", as used herein, is defined as that which is produced by *in vitro* chemical or enzymatic synthesis. The synthetic arrays of the present invention may be contrasted with natural nucleic acid molecules such as viral or plasmid vectors, for instance, which may be propagated in bacterial, yeast, or other living hosts.

As used herein, the term "nucleic acid" is defined to encompass DNA and RNA or both synthetic and natural origin. The nucleic acid may exist as single- or double-stranded DNA or RNA, an RNA/DNA heteroduplex or an RNA/DNA copolymer, wherein the term "copolymer" refers to a single nucleic acid strand that comprises both ribonucleotides and deoxyribonucleotides.

The phrase "different nucleic acid molecule members" means that the double-stranded nucleic acid molecules attached to the surface include double-stranded nucleic acid molecules of of different nucleotide sequence.

When used herein in this context, the term "double-stranded" refers to a pair of nucleic acid molecules, as defined above, that exist in a hydrogen-bonded, helical array typically associated with DNA, and that under these umbrella terms are included those paired oligonucleotides that are essentially double-stranded, meaning those that contain short regions of mismatch, such as a mono-, di- or tri-nucleotide, resulting from design or error either in chemical synthesis of the oligonucleotide priming site on the first nucleic acid strand or in enzymatic synthesis of the second nucleic acid strand.

As used herein, the terms "complementary" and "substantially complementary" refer to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double-stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single-stranded RNA or DNA molecules are said to be substantially complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 80% of the nucleotides

of the other strand, usually at least about 90% to 95%, and more preferably from about 98 to 100%.

As used herein, the term "array" is defined to mean a heterogeneous pool of nucleic acid molecules that is affixed to a substrate or solid support in a manner that permits identification of individual members during the course of experimental manipulation.

According to the invention, the array may have virtually any number of different members. In preferred embodiments, the array comprises from 2 up to 100 members, more preferably from 100 up to 10,000 members and highly preferably from 10,000 up to 1,000,000 members, preferably on a solid support. In preferred embodiments, the array will have a density of more than 100 members at known locations per cm^2, preferably more than 1,000 per cm^2, more preferably more than 10,000 per cm^2.

According to the methods disclosed herein, a "substrate" or "solid support" is defined as any material having a rigid or semi-rigid surface.

It is contemplated that attached to the solid support is a spacer. The spacer molecule is preferably of sufficient length to permit the double-stranded oligonucleotide in the completed member of the array to interact freely with molecules exposed to the array. The spacer molecule, which may comprise as little as a covalent bond length, is typically 6-50 atoms long to provide sufficient exposure for the attached double-stranded DNA molecule. The spacer is comprised of a surface attaching portion and a longer chain portion.

Preferably, the 3' end of the first strand is linked to the solid support.

It is preferred that the 5' end of the first strand and 3' end of the second strand are not linked *via* a covalent bond, and thus do not form a continuous single strand. As used herein in this context, "covalent bond" is defined as meaning a bond that forms, directly or *via* a spacer comprising nucleic acid or another material, a continuous strand that comprises the 5' end of the first strand and the 3' end of the second strand, and thus includes a 3'/5' phosphate bond as occurs naturally in a single-stranded nucleic acid. This definition does not encompass intermolecular crosslinking of the first and second strands.

It is additionally preferred that the 5' end of the second strand is not linked to the support.

It is preferred that the solid support is a silica support.

It is also preferred that the first strand is produced by chemical synthesis and that the second strand is produced by enzymatic synthesis.

Preferably, the first strand is used as the template on which the second strand is enzymatically produced.

It is additionally preferred that in each member of the array, the first strand contains at its 3' end a binding site for an oligonucleotide primer which is used to prime enzymatic synthesis of the second strand, and at its 5' end a variable sequence.

An "oligonucleotide primer", as referred to herein, is defined as a single-stranded DNA or RNA molecule that is hybridized to a nucleic acid template to prime enzymatic synthesis of a second nucleic acid strand.

It is preferred that enzymatic synthesis of the second strand is performed using an enzyme. Preferably, the oligonucleotide primer is between 10 and 30 nucleotides in length.

It is preferred that the first strand comprises DNA.

It is additionally preferred that the second strand comprises DNA.

It is preferred that the first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

Use of the term "monomer" is made to indicate any of the set of molecules which can be joined together to form an oligomer or polymer. The set of monomers useful in the present invention includes, but is not restricted to, for the example of oligonucleotide synthesis, the set of nucleotides consisting of adenine, thymine, cytosine, guanine, and uridine (A, T, C, G, and U, respectively) and synthetic analogs thereof. As used herein, "monomer" refers to any member of a basis set for synthesis of an oligomer. Different basis sets of monomers may be used at successive steps in the synthesis of a polymer.

In a particularly preferred embodiment, the solid support is a silica support and the first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

It is also preferred that a chimeric protein comprising a DNA binding domain fused in-frame to Green Fluorescent Protein is bound to nucleic acid molecules of said array.

The present invention also provides a method for the construction of a synthetic, surface-bound nucleic acid array, comprising the steps of

(a) performing chemical synthesis of a first nucleic acid strand that is linked to a solid support, (b) hybridizing to the first strand of step (a) an oligonucleotide primer that is substantially complementary to a sequence comprised by the first strand, and (c)

performing enzymatic synthesis of a second nucleic acid strand that is complementary to the first strand of step (a), wherein the second strand is complexed to the first strand by Watson-Crick base pairing.

Preferably, the 3' end of the first strand is linked to the solid support.

It is preferred that the 5' end of the first strand and the 3' end of the second strand are not linked *via* a covalent bond.

It is additionally preferred that the 5' end of the second strand is not linked to the support.

Preferably, the solid support is a silica support.

It is also preferred that in each member of the array, the first strand contains at its 3' end a binding site for an oligonucleotide primer which is used to prime enzymatic synthesis of the second strand, and at its 5' end a variable sequence.

It is additionally preferred that the enzymatic synthesis of the second strand is performed using an enzyme.

Preferably, the oligonucleotide primer of step (c) is between 10 and 30 nucleotides in length.

In a preferred embodiment, the first strand comprises DNA.

It is additionally preferred that the second strand comprises DNA.

Preferably, the first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

Preferably, the solid support is a silica support. More preferably, the solid support is a silica support and the first and second strands each comprise from 16 to 60 monomers selected f rom the group that includes ribonucleotides and deoxyribonucleotides.

The invention provides an improvement over know nucleic acid arrays or libraries in that while the first strand of the DNA duplex is chemically-synthesized on the support matrix, the second strand is enzymatically produced using the first strand as a template. While the error rate in production of the first strand remains the same, increased fidelity of second strand synthesis is expected, consequently, to result in a higher percentage of points on the matrix surface that are filled by hybridized DNA duplex molecules that can serve as targets for binding- or other assays. In addition, oligonucleotide priming of second strand synthesis obviates the need for covalent linkage of complementary regions, with the effect of reducing extraneous sequence or non-nucleic acid material from the array, as well as eliminating steps

of designing and synthesizing such a linker.

Further features and advantages of the invention will become more fully apparent in the following description of the embodiments and drawings thereof, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 presents a schematic summary of light-directed DNA synthesis.

Figure 2 presents a photomicrograph of a fluorescently-labeled array of bimolecular, double-stranded DNA molecules on a silica chip.

Figure 3 presents confocal argon laser scanning to detect fluorescently-labeled, surface-bound nucleic acid molecules.

Figure 4 presents *Rsa*I digestion of a fluorescently-labeled array of bimolecular, double-stranded DNA molecules on a silica chip.

Figure 5 presents binding of Green Fluorescent Protein to an array of bimolecular, double-stranded DNA molecules on a silica chip, and confocal argon laser scanning to detect the bound protein.

DESCRIPTION

Bimolecular Double-Stranded Arrays According To The Invention

The invention is based on the recognition that bimolecular double-stranded nucleic acid molecule arrays may be provided, and that such arrays possess the advantage of a high fidelity of second strand synthesis, and are therefore provide an array of true duplex nucleic acid. Described below is how to prepare an array of immobilized first strand, how to prepare and/or design a primer useful according to the invention, and how to primer synthesis of a second strand that is complementary to and duplexed with the first array-bound strand.

Preparation of Array of Immobilized First DNA Strand

Synthesis of the nucleic acid arrays of the present invention is a bipartite process, which entails the production of a diverse array of single-stranded DNA molecules that are immobilized on a the surface of a solid support matrix, followed by priming and enzymatic synthesis of a second nucleic acid strand, either RNA or DNA. A highly preferred method of carrying out synthesis of the immobilized single-stranded array is that of Lockhart, described

in U.S. Patent No. 5,556,752 the contents of which are herein incorporated by reference. Of the methods described therein, that which is of particular use describes synthesis of such an array on the surface of a single solid support has a plurality of preselected regions. A method whereby each chemically distinct member of the array is synthesized on a separate solid support is also described by Lockhart. These methods, and others, are briefly summarized below.

The solid support may comprise biological, nonbiological, organic or inorganic materials, or a combination of any of these. It is contemplated that such materials may exist as particles, strands, precipitates, gels, sheets, tubing, spheres, containers, capillaries, pads, slices, films, plates or slides. Preferably the solid support takes the form of plates or slides, small beads, pellets, disks or other convenient forms. It is highly preferred that at least one surface of the substrate will be substantially flat. The solid support may take on alternative surface configurations. For example, the solid support may contain raised or depressed regions on which synthesis takes place. In some instances, the solid support will be chosen to provide appropriate light-absorbing characteristics. For example, the support may be a polymerized Langmuir Blodgett film, functionalized glass, Si, Ge, GaAs, GaP, SiO2, SiN4, modified silicon, or any one of a variety of gels or polymers such as (poly)tetrafluoroethylene, (poly)vinylidendifluoride, polystyrene, polycarbonate, or combinations thereof. Other suitable solid support materials may be used, and will be readily apparent to those of skill in the art. Preferably, the surface of the solid support will contain reactive groups, which could be carboxyl, amino, hydroxyl, thiol, or the like. More preferably, the surface will be optically transparent and will have surface Si-OH functionalities, such as are found on silica surfaces.

According to the invention, a first nucleic acid strand is anchored to the solid support by as little as an intermolecular covalent bond. Alternatively, a more elaborate linking molecule may attach the nucleic acid strand to the support. Such a molecular tether may comprise a surface-attaching portion which is directly attached to the solid support. This portion can be bound to the solid support *via* carbon-carbon bonds using, for example, supports having (poly)trifluorochloroethylene surfaces, or preferably, by siloxane bonds (using, for example, glass or silicon oxide as the solid support). Siloxane bonds with the surface of the support can be formed *via* reactions of surface attaching portions bearing trichlorosilyl or trialkoxysilyl groups. The surface attaching groups will also have a site for

attachment of the longer chain portion. It is contemplated that suitable attachment groups may include amines, hydroxyl, thiol, and carboxyl groups. Preferred surface attaching portions include aminoalkylsilanes and hydroxyalkylsilanes. It is particularly preferred that the surface attaching portion of the spacer is selected from the group comprising bis(2-hydroxyethyl)-aminopropyltriethoxysilane, 2-hydroxyethylaminopropyltriethoxysilane, aminopropyltriethoxysilane and hydroxypropyltriethoxysilane.

The longer chain portion of the spacer can be any of a variety of molecules which are inert to the subsequent conditions for polymer synthesis, examples of which include: aryl acetylene, ethylene glycol oligomers containing 2-14 monomer units, diamines, diacids, amino acids, peptides, or combinations thereof. It is contemplated that the longer chain portion is a polynucleotide. The longer chain portion which is to be used as part of the spacer can be selected based upon its hydrophilic/hydrophobic properties to improve presentation of the double-stranded oligonucleotides to certain receptors, proteins or drugs. It can be constructed of polyethyleneglycols, polynucleotides, alkylene, polyalcohol, polyester, polyamine, polyphosphodiester and combinations thereof.

Additionally, for use in synthesis of the arrays of the invention, the spacer will typically have a protecting group, attached to a functional group (i.e., hydroxyl, amino or carboxylic acid) on the distal or terminal end of the chain portion (opposite the solid support). After deprotection and coupling, the distal end is covalently bound to an oligomer.

As used in discussion of the spacer region, the term "alkyl" refers to a saturated hydrocarbon radical which may be straight -chain or branced-chain (for example, ethyl,isopropyl, t-amyl, or 2,5-0dimethylhexyl). When "alkyl" or "alkylene" is used to refer to a linking group or a spacer, it is taken to be a group having two available valences for covalent attachment, for example, --CH_2CH_2--, --$CH_2CH_2CH_2$--, -$CH_2CH_2CH(CH_3)CH_2$-- -- $CH_2(CH_2CH_2)_2CH_2$--. Preferred alkyl groups as substitutents are those containing 1 to 10 carbon atoms, with those containing 1 ato 6 carbon atoms being particularly preferred. Preferred alkyl or alkylene groups as linking groups are those containing 1 to 20 carbon atoms, with those containing 3 to 6 carbon atoms being particularly preferred. The term "polyethylene glycol" is used to refer to those molecules which have repeating units of ethylene glycol, for example, hexaethylene glycol (HO--$(CH_2CH_20)_5$--$CH_2(CH_2CH_2OH)$. When the term "polyethylene glycol" is used to refer to linking groups and spacer groups, it would be understood by one of skill in the art that other polyethers of polyols could be used

as well (i.e., polypropylene glycol or mistures of ethylene and propeylene glycols).

The term "protecting group", as used herein, refers to any of the groups which are designed to block one reactive site in a molecule while a chemical reaction is carried out at another reactive site. More particularly, the protecting groups used herein can be any of those groups described in Greene et al., 1991, Protective Groups In Organic Chemistry, 2nd Ed., John Wiley & Sons, New York, N.Y, incorporated herein by reference. The proper selection of protecting groups for a particular synthesis will be governed by the overall methods employed in the synthesis. For example, in "light-directed" synthesis, discussed below, the protecting groups will be photolabile protecting groups, e.g. NVOC and MeNPOC. In other methods, protecting groups may be removed by chemical methods and include groups such as FMOC, DMT and others known to those of skill in the art.

Nucleic Acid Arrays on a Single Substrate

1. Light-directed methods

Where a single solid support is employed, the oligonucleotides of the present invention can be formed using a variety of techniques known to those skilled in the art of polymer synthesis on solid supports. For example, "light-directed" methods, techniques in a family of methods known as VLSIPS™ methods, are described in U.S. Patent No. 5,143,854 and U.S. Patent No. 5,510,270 and U.S. Patent No. 5,527,681, which are herein incorporated by reference. These methods, which are illustrated in Figure 1 (adapted from Pease et al., 1994, Proc. Natl. Acad. Sci. U.S.A., 91: 5022-5026), involve activating predefined regions of a substrate or solid support and then contacting the substrate with a preselected monomer solution. These regions can be activated with a light source, typically shown through a mask (much in the manner of photolithography techniques used in integrated circuit fabrication). Other regions of the substrate remain inactive because illumination is blocked by the mask and they remain chemically protected. Thus, a light pattern defines which regions of the substrate react with a given monomer. By repeatedly activating different sets of predefined regions and contacting different monomer solutions with the substrate, a diverse array of polymers is produced on the substrate. Other steps, such as washing unreacted monomer solution from the substrate, can be used as necessary. Other applicable methods include mechanical techniques such as those described in PCT No. 92/10183, U.S. Pat. No. 5,384,261 also incorporated herein by reference for all purposes. Still further techniques include bead based techniques such as those described in PCT US/93/04145, also incorporated herein by

reference, and pin based methods such as those described in U.S. Pat. No. 5,288,514, also incorporated herein by reference.

The VLSIPS™ methods are preferred for making the compounds and arrays of the present invention. The surface of a solid support, optionally modified with spacers having photolabile protecting groups such as NVOC and MeNPOC, is illuminated through a photolithographic mask, yielding reactive groups (typically hydroxyl groups) in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at the 5'-hydroxyl with a photolabile protecting group) is then presented to the surface and chemical coupling occurs at sites that were exposed to light. Following capping and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of oligonucleotides is produced. Alternatively, an oligomer of from, for example, 4 to 30 nucleotides can be added to each of the preselected regions rather than synthesize each member in one nucleotide monomer at a time.

2. Flow Channel or Spotting Methods

Additional methods applicable to array synthesis on a single substrate are described in U.S. Patent No. 5,384,261, incorporated herein by reference for all purposes. In the methods disclosed in these applications, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions. Other approaches, as well as combinations of spotting and flowing, may be employed as well. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to arrays of the present invention can generally be described as follows: Diverse polymer sequences are synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A

flows through or is placed in all or some of the channel(s). The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in fluid contact with a second flow channel(s) through translation, rotation, or replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter, the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed through still other channels, etc. In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

The "spotting" methods of preparing compounds and arrays of the present invention can be implemented in much the same manner. A first monomer, A, can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a second monomer, B, can be delivered to and reacted with a second group of activated reaction regions. Unlike the flow channel embodiments described above, reactants are delivered in relatively small quantities by directly depositing them in selected regions. In

some steps, the entire substrate surface can be sprayed or otherwise coated with a solution, if it is more efficient to do so. Precisely measured aliquots of monomer solutions may be deposited dropwise by a dispenser that moves from region to region. Typical dispensers include a micropipette to deliver the monomer solution to the substrate and a robotic system to control the position of the micropipette with respect to the substrate, or an ink-jet printer. In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes, or the like so that various reagents can be delivered to the reaction regions simultaneously.

3. Pin-Based Methods

Another method which is useful for the preparation of the immobilized arrays of single-stranded DNA molecules X of the present invention involves "pin-based synthesis." This method, which is described in detail in U.S. Patent No. 5,288,514, previously incorporated herein by reference, utilizes a substrate having a plurality of pins or other extensions. The pins are each inserted simultaneously into individual reagent containers in a tray. An array of 96 pins is commonly utilized with a 96-container tray, such as a 96-well microtitre dish.

Each tray is filled with a particular reagent for coupling in a particular chemical reaction on an individual pin. Accordingly, the trays will often contain different reagents. Since the chemical reactions have been optimized such that each of the reactions can be performed under a relatively similar set of reaction conditions, it becomes possible to conduct multiple chemical coupling steps simultaneously. The invention provides for the use of substrate(s) on which the chemical coupling steps are conducted. The substrate is optionally provided with a spacer, S, having active sites. In the particular case of oligonucleotides, for example, the spacer may be selected from a wide variety of molecules which can be used in organic environments associated with synthesis as well as aqueous environments associated with binding studies such as may be conducted between the nucleic acid members of the array and other molecules. These molecules include, but are not limited to, proteins (or fragments thereof), lipids, carbohydrates, proteoglycans and nucleic acid molecules. Examples of suitable spacers are polyethyleneglycols, dicarboxylic acids, polyamines and alkylenes, substituted with, for example, methoxy and ethoxy groups. Additionally, the spacers will have an active site on the distal end. The active sites are optionally protected initially by protecting groups. Among a wide variety of protecting groups which are useful

are FMOC, BOC, t-butyl esters, t-butyl ethers, and the like.

Various exemplary protecting groups are described in, for example, Atherton et al., 1989, Solid Phase Peptide Synthesis, IRL Press, incorporated herein by reference. In some embodiments, the spacer may provide for a cleavable function by way of, for example, exposure to acid or base.

Arrays on Multiple Substrates

Yet another method which is useful for synthesis of compounds and arrays of the present invention involves "bead based synthesis." A general approach for bead based synthesis is described in PCT/US93/04145 (filed Apr. 28, 1993), the disclosure of which is incorporated herein by reference.

For the synthesis of molecules such as oligonucleotides on beads, a large plurality of beads are suspended in a suitable carrier (such as water) in a container. The beads are provided with optional spacer molecules having an active site to which is complexed, optionally, a protecting group.

At each step of the synthesis, the beads are divided for coupling into a plurality of containers. After the nascent oligonucleotide chains are deprotected, a different monomer solution is added to each container, so that on all beads in a given container, the same nucleotide addition reaction occurs. The beads are then washed of excess reagents, pooled in a single container, mixed and re-distributed into another plurality of containers in preparation for the next round of synthesis. It should be noted that by virtue of the large number of beads utilized at the outset, there will similarly be a large number of beads randomly dispersed in the container, each having a unique oligonucleotide sequence synthesized on a surface thereof after numerous rounds of randomized addition of bases. As pointed out by Lockhart (U.S. Patent No. 5,556,752) an individual bead may be tagged with a sequence which is unique to the double-stranded oligonucleotide thereon, to allow for identification during use.

Preparation Of Oligonucleotide Primers Useful In The Invention

Oligonucleotide primers useful according to the invention are single-stranded DNA or RNA molecules that are hybridizable to a nucleic acid template to prime enzymatic synthesis of a second nucleic acid strand. The primer may therefore be of any sequence composition or length, provided it is complementary to a portion of the first strand.

It is contemplated that such a molecule is prepared by synthetic methods, either

chemical or enzymatic. Alternatively, such a molecule or a fragment thereof may be naturally occurring, and may be isolated from its natural source or purchased from a commercial supplier. It is contemplated that oligonucleotide primers employed in the present invention will be 6 to 100 nucleotides in length, preferably from 10 to 30 nucleotides, although oligonucleotides of different length may be appropriate.

Additional considerations with respect to design of a selected primer useful according to the invention relate to duplex formation, and are described in detail in the following section.

Oligonucleotide Primer Hybridization To Single-stranded Nucleic Acid Sequences And Extension to Form Bimolecular Double-Stranded Nucleic Acids

Of central importance in carrying out the method of the present invention is selective hybridization of an oligonucleotide primer to the first nucleic acid strand in order to permit enzymatic synthesis of the second nucleic acid strand. Any of a number of enzymes well known in the art can be utilized in the synthesis reaction. Preferably, enzymatic synthesis of the second strand is performed using an enzyme selected from the group comprising DNA polymerase I ($exo^{(-)}$ Klenow fragment), T4 DNA polymerase, T7 DNA polymerase, modified T7 DNA polymerase, Taq DNA polymerase, $exo^{(-)}$ vent DNA polymerase, $exo^{(-)}$ deep vent DNA polymerase, reverse transcriptase and RNA polymerase.

Typically, selective hybridization will occur when two nucleic acid sequences are substantially complementary (typically, at least about 65% complementary over a stretch of at least 14 to 25 nucleotides, preferably at least about 75%, more preferably at least about 90% complementary). See Kanehisa, M., 1984, Nucleic Acids Res. 12: 203, incorporated herein by reference. As a result, it is expected that a certain degree of mismatch at the priming site can be tolerated. Such mismatch may be small, such as a mono-, di- or tri-nucleotide. Alternatively, it may encompass loops, which we define as regions in which mismatch encompasses an uninterrupted series of four or more nucleotides. Note that such loops within the oligonucleotide priming site are encompassed by the present invention; however, the invention does not provide double-stranded nucleic acids that comprise loop structures between the 5' end of the first strand and the 3' end of the second strand. In addition, loop structures outside the priming site, but which do not encumber the 5' end of the first strand or the 3' end of the second strand are not provided by the present invention, since there is no

known mechanism for generating such structures in the course of enzymatic second-strand nucleic acid synthesis. Both the 5' end of the first strand and the 3' end of the second strand must be free of attachment to each other via a continuous single strand.

Either strand may comprise RNA or DNA. Overall, five factors influence the efficiency and selectivity of hybridization of the primer to the immobilized first strand. These factors are (i) primer length, (ii) the nucleotide sequence and/or composition, (iii) hybridization temperature, (iv) buffer chemistry and (v) the potential for steric hindrance in the region to which the probe is required to hybridize.

There is a positive correlation between primer length and both the efficiency and accuracy with which a primer will anneal to a target sequence; longer sequences have a higher T_M than do shorter ones, and are less likely to be repeated within a given first nucleic acid strand, thereby cutting down on promiscuous hybridization. Primer sequences with a high G-C content or that comprise palindromic sequences tend to self-hybridize, as do their intended target sites, since unimolecular, rather than bimolecular, hybridization kinetics are genererally favored in solution; at the same time, it is important to design a primer containing sufficient numbers of G-C nucleotide pairings to bind the target sequence tightly, since each such pair is bound by three hydrogen bonds, rather than the two that are found when A and T bases pair. Hybridization temperature varies inversely with primer annealing efficiency, as does the concentration of organic solvents, e.g. formamide, that might be included in a hybridization mixture, while increases in salt concentration facilitate binding. Under stringent hybridization conditions, longer probes must be used, while shorter ones will suffice under more permissive conditions. Stringent hybridization conditions will typically include salt concentrations of less than about 1M, more usually less than about 500 mM and preferably less than about 200 mM. Hybridization temperatures can be as low as 5°C, but are typically greater than 22°C, more typically greater than about 30°C, and preferably in excess of about 37°C. Longer fragments may require higher hybridization temperatures for specific hybridization. As several factors may affect the stringency of hybridization, the combination of parameters is more important than the absolute measure of any one alone.

Primers must be designed with the above first four considerations in mind. While estimates of the relative merits of numerous sequences can be made mentally, computer programs have been designed to assist in the evaluation of these several parameters and the optimization of primer sequences. Examples of such programs are "PrimerSelect" of

DNAStar and OLIGO™. Once designed, suitable oligonucleotides may be prepared by the phosphoramidite method described by Beaucage and Carruthers, 1981, Tetrahedron Lett., 22: 1859-1862, or by the triester method according to Matteucci et al., 1981, J. Am. Chem. Soc., 103: 3185, both incorporated herein by reference, or by other chemical methods using either a commercial automated oligonucleotide synthesizer or VLSIPS™ technology (discussed in detail below).

The fifth consideration, steric hindrance, is one that was of particular relevance to the development of the invention disclosed herein. While methods for the primed, enzymatic synthesis of second nucleic acid strands from immobilized first strands are known in the art (see Uhlen, U.S. Patent No. 5,405,746 and Utermohlen, U.S. Patent No. 5,437,976), the present method differs in that the priming site, as determined by the location of the 3' end of the first strand (X), is adjacent to the surface of the solid support. In a typical silica-based chip array, made as per Lockhart (U.S. Patent No. 5,556,752), a 20 μm^2 region carries approximates 4×10^6 functional copies of a specific sequence, with an intermolecular spacing distance of about 100 Å (Chee et al., 1996, Science, 274: 610-614). As a result, it is necessary that the oligonucleotide primer hybridize *efficiently* to an anchored target in a confined space, and that synthesis proceed outward from the support. In the above-referenced disclosures, it is the 5' end of the first oligonucleotide strand which is linked to the matrix; therefore, priming of the free end of that molecule is permitted, and second-strand extension proceeds toward the solid support. Under the circumstances, significant uncertainty existed as to whether oligonucleotide priming of the end of the first strand proximal to the solid support would occur at a sufficiently high frequency to yield a high-density double-stranded nucleic acid array. The surprising success of this method is described below in Example 1.

EXAMPLE 1

This example illustrates the general synthesis of an array of bimolecular, double-stranded oligonucleotides on a solid support.

As a first step, single-stranded DNA molecules were synthesized on a solid support using standard light-directed methods (VLSIPS™ protocols), as as described above, using the method of Lockhart, U.S. Patent No. 5,556,752, the contents of which incoporated above by reference. Hexaethylene glycol (PEG) linkers were used to covalently attach the synthesized oligonucleotides to the derivatized glass surface. A heterogeneous array of linkers was formed such that some sectors of the silica chip had linkers comprising two PEG linkers,

while other sectors bore linkers comprising a single PEG molecule (Figure 2). In addition, the intermolecular distance between linker molecules (and, consequently, nascent nucleic acid strands) was varied such that for either length of linker and for each of the 9,600 distinct molecular species synthesized, were 15 different chip sectors representing the following range of strand densities. These densities, expressed as the percent of total anchoring sites occupied by nucleic acid molecules, are shown in Table 1.

Table 1

% of sites filled	% of sites filled, cont'd.	% of sites filled, cont'd.
0.4	25.0	69.1
1.6	31.5	75.8
3.1	39.7	83.1
6.2	50.0	91.2
12.5	63.0	100.0

Synthesis of the first strand proceeded one nucleotide at a time using repeated cycles of photo-deprotection and chemical coupling of protected nucleotides. The nucleotides each had a protecting group on the base portion of the monomer as well as a photolabile MeNPoc protecting group on the 5' hydroxyl. Note that each of the different molecular species occupies a different physical region on the chip so that there is a one-to-one correspondence between molecular identity and physical location. Moving outward from the chip, the sequence of each molecule proceeds from its 3' to its 5' end (the 3' end of the DNA molecule is attached to the solid surface via a silyl group and 2 PEG linkers), as is the case when chemical synthetic methods are utilized.

Second strand synthesis, as stated above, requires priming of a site at the 3' end of the first nucleic acid strand, followed by enzymatic extension of the primed sequence. DNA polymerase I ($exo^{(-)}$ Klenow fragment) was employed in this experiment, although numerous other enzymes, as discussed above, may be advantageously employed. This particular enzyme is optimally active at 37°C; therefore, two priming sites and the corresponding complementary primers were designed that were predicted to bind efficiently and yet exhibit a minimum of secondary structure at that temperature according to calculations performed by the DNAStar "PrimerSelect" computer program, which was employed for this purpose. The

sequences of these primers are as follows:

1s 5'--TCCACACTCTCCAACA--3' [SEQ ID NO: 1] (estimated T_M = 36.8°C)

2s 5'--GGACCCTTTGACTTGA--3' [SEQ ID NO: 2] (estimated T_M = 38.7°C)

Note that the optimal reaction temperature varies considerably among polymerases. Also of use according to the methods of the invention are exo$^{(-)}$ vent DNA polymerase and exo$^{(-)}$ deep vent DNA polymerase (both commercially available from New England Biolabs, Beverly, MA), which are optimally active at 72°C and approximately 30% active at 50°C, according to the manufacturer. Were these enzymes to be used, longer primer sequences, or those with a higher G-C content, would have to be employed.

In the case of the synthesis presented in Figure 2, primer S1 [SEQ ID NO: 1] was used. The reaction conditions were as follows:

Prehybridization of chip: 0.005% Triton X-100, 0.2 mg/ml acetylated bovine serum albumin (BSA), 10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$ and 7.5 mM dithiothreitol (DTT) at 37°C for 30 to 60 minutes on a rotisserie.

Second-strand primer extension and fluorescein labeling: 0.005% Triton, 10 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, 7.5 mM DTT, 0.4 mM dNTP's, 0.4 μM primer, 0.04 U/μl DNA Polymerase I (3' to 5' exo$^{(-)}$ Klenow fragment, New England Biolabs, Beverly, MA) and 0.0004 mM of fluorescein-12-labeled dATP at 37°C on a for 1 to 2 hours on a rotisserie, followed by a wash in 0.005% Triton X-100 in 6× SSPE at room temperature. (Note that an alternate labeling procedure, not used in the experiment presented in this Example, is one in which unlabeled extension is performed, followed by labeled primer extension using terminal deoxynucleotide transferase. This reaction takes place as follows: 0.005% Triton X-100, 10 mM Tris acetate, pH 7.5, 10 mM magnesium acetate, 50 mM potassium acetate, 0.044 U/μl terminal transferase and 0.014 mM of any fluorescein-12-labeled dideoxynucleotide at 37°C for 1-2 hr. on a rotisserie, followed by a wash in 0.005% Triton X-100 in 6× SSPE at room temperature.)

To confirm that second-strand synthesis had taken place, the chip was scanned under a layer of wash buffer for fluorescence in an argon laser confocal scanner (see U.S. Patent No. 5,578,832). This device exposes the molecules of the array to irradiation at a wavelength of 488 nanometers, which excites electrons in the fluorescein moiety, resulting in fluorescent

emissions, which are then recorded at each position of the chip (Figure 3). Since the first strand was unlabeled, the efficiency of second-strand synthesis can be measured. The result is shown in Figure 2, where various sectors of the chip fluoresce with different intensities, in proportion both to strand density and to the proportion of dATP residues in the second strand.

Further confirmation of successful second-strand synthesis was gained from a biochemical assay of the chip. According to the first-strand synthesis procedure, several sectors of the chip were designed such that the several unique sequences synthesized at those positions contained a 4 base motif which, when double-stranded, would form an endonuclease recognition site for the enzyme *Rsa*I. The chip was digested in *Rsa*I, using the manufacturer's recommended incubation conditions. Upon re-scanning of the chip in the argon laser scanner, a dark area appeared. This can be seen in Figure 2, and is shown in detail in Figure 4. Since the ability of the enzyme to cleave the sequence from the chip is dependent upon the sequence being double-stranded, synthesis, at least to the point of the recognition site, must have occurred.

In addition to providing evidence of successful second-strand synthesis, cleavage of double-stranded nucleic acid molecules from the solid support with *Rsa*I demonstrates that members of the array are accessible to proteins in solution, a requirement if the arrays of the invention are to be useful in carrying out assays of protein/DNA interactions. We have devised a procedure in which chimeric proteins, each comprising a DNA binding domain fused in-frame to Green Fluorescent Protein (GFP), are incubated with arrays produced according to the methods of the invention in order to determine a consensus sequence for a given protein/DNA binding motif. After washing away any unbound fusion protein, the support bearing the array is scanned with the confocal argon laser (Figure 5); the intensity of fluorescence, which is proportional to the amount of protein bound, is correlated with the sequences of nucleic acid molecules, which are known at each position of the scanned surface. The range of sequences to which a protein will bind, as well as the relative efficiency of binding to each, can then be determined. In order to interpret the results, the only source of fluorescence on the chip must be GFP; therefore, the nucleic acid molecules of the array must be unlabeled. The strand extension reaction described above can, if desired, be performed without the use of a fluorescent label; the reaction conditions are identical except that the fluorescein-labeled dATP is omitted, along with the wash step, the purpose of which is to remove unincorporated background fluorescence that ordinarily might interfere with

scanning.

USE

The present invention is useful for the production of accurate, high-density arrays of double-stranded nucleic acid molecules the sequences of which can be determined based upon physical location within the array. The arrays provided are, themselves, useful as substrates for multiplex screening of compounds that might physically interact with such nucleic acid sequences, such as DNA binding proteins or other compositions that are of potential scientific or clinical interest, particularly those with therapeutic potential.

OTHER EMBODIMENTS

Other embodiments will be evident to those of skill in the art. It should be understood that the foregoing description is provided for clarity only and is merely exemplary. The spirit and scope of the present invention are not limited to the above examples, but are encompassed by the following claims.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Church, George M.

Bulyk, Martha L.

(ii) TITLE OF INVENTION: SURFACE-BOUND, BIMOLECULAR, DOUBLE-STRANDED DNA ARRAYS

(iii) NUMBER OF SEQUENCES: 2

(iv) CORRESPONDENCE ADDRESS:

(A) ADDRESSEE: Banner & Witcoff, Ltd.

(B) STREET: 75 State Street

(C) CITY: Boston

(D) STATE: MA

(E) COUNTRY: US

(F) ZIP: 02109

(v) COMPUTER READABLE FORM:

(A) MEDIUM TYPE: Diskette, 3.50 inch, 1.44 Mb storage

(B) COMPUTER: IBM PC compatible

(C) OPERATING SYSTEM: PC-DOS/MS-DOS

(D) SOFTWARE: WordPerfect 6.1

(vi) CURRENT APPLICATION DATA:

APPLICATION NUMBER:

FILING DATE: 05-AUG-1997

(vii) PRIOR APPLICATION DATA:

(A) APPLICATION:

(B) FILING DATE:

(viii) ATTORNEY/AGENT INFORMATION:

(A) NAME: Williams, Ph.D., Kathleen M.

(B) REGISTRATION NUMBER: 34,380

(C) REFERENCE/DOCKET NUMBER: 10498/10932

(ix) TELECOMMUNICATION INFORMATION:

(A) TELEPHONE: (617) 345-9100

(B) TELEFAX: (617) 345-9111

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 bases

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

TCCACACTCT CCAACA 16

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 bases

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

GGACCCTTTG ACTTGA 16

CLAIMS

What is claimed is:

1. A synthetic array of surface-bound, bimolecular, double-stranded, nucleic acid molecules, said array comprising

a solid support, and

a plurality of different double-stranded nucleic acid molecule members, a said member comprising a first nucleic acid strand linked to said solid support and a second nucleic acid strand which is substantially complementary to said first strand and complexed to said first strand by Watson-Crick base pairing, wherein at least a portion of said members have a second nucleic acid strand is substantially complementary to and base paired with said first strand along the entire length of said first strand.

2. The array of claim 1, wherein the 3' end of said first strand is linked to said support.

3. The array of claim 1, wherein the 5' end of said first strand and the 3' end of said second strand are not linked *via* a covalent bond.

4. The array of claim 1, wherein the 5' end of said second strand is not linked to said support.

5. The array of claim 1, wherein said solid support is a silica support.

6. The array of claim 1, wherein said first strand is produced by chemical synthesis and said second strand is produced by enzymatic synthesis.

7. The array of claim 6, wherein said first strand is used as the template on which said second strand is enzymatically produced.

8. The array of claim 7, wherein said first strand of each member of said array contains at its 3' end a binding site for an oligonucleotide primer which is used to prime enzymatic synthesis of said second strand, and at its 5' end a variable sequence.

9. The array of claim 6, wherein said enzymatic synthesis is performed using an enzyme.

10. The array of claim 8, wherein said oligonucleotide primer is between 10 and 30 nucleotides in length.

11. The array of claim 1, wherein said first strand comprises DNA.

12. The array of claim 1, wherein said second strand comprises DNA.

13. The array of claim 1, wherein said first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

14. The array of claim 1, wherein said solid support is a silica support and said first and second strands X each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

15. The array of claim 1, wherein a chimeric protein comprising a DNA binding domain fused in-frame to Green Fluorescent Protein is bound to nucleic acid molecules of said array.

16. A method for the construction of a synthetic, surface-bound nucleic acid array, comprising the steps of

(a) providing a first nucleic acid strand that is linked to a solid support,

(b) hybridizing to said first strand of step (a) an oligonucleotide primer that is substantially complementary to a sequence comprised by said first strand, and

(c) performing enzymatic synthesis of a second nucleic acid strand that is complementary to said first strand of step (a), wherein said second strand is complexed to said first strand by Watson-Crick base pairing.

17. The method according to claim 16, wherein the 3' end of said first strand is linked to said support.

18. The method according to claim 16, wherein the 5' end of said first strand and the 3' end of said second strand are not linked *via* a covalent bond.

19. The method according to claim 16, wherein the 5' end of said second strand is not linked to said solid support.

20. The method according to claim 16, wherein said solid support is a silica support.

21. The method according to claim 16, wherein said first strand of each member of said array contains at its 3' end a binding site for an oligonucleotide primer which is used to prime enzymatic synthesis of said second, and at its 5' end a variable sequence, wherein said binding site is present in each said member of said array.

22. The method according to claim 16, wherein said enzymatic synthesis is performed using an enzyme.

23. The method according to claim 16, wherein said oligonucleotide primer of step (b) is between 10 and 30 nucleotides in length.

24. The method according to claim 16, wherein said first strand of step (a) comprises DNA.

25. The method according to claim 16, wherein said second strand of step (c) comprises DNA.

26. The method according to claim 16, wherein said first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

27. The method according to claim 16, wherein said solid support is a silica support and said first and second strands each comprise from 16 to 60 monomers selected from the group that includes ribonucleotides and deoxyribonucleotides.

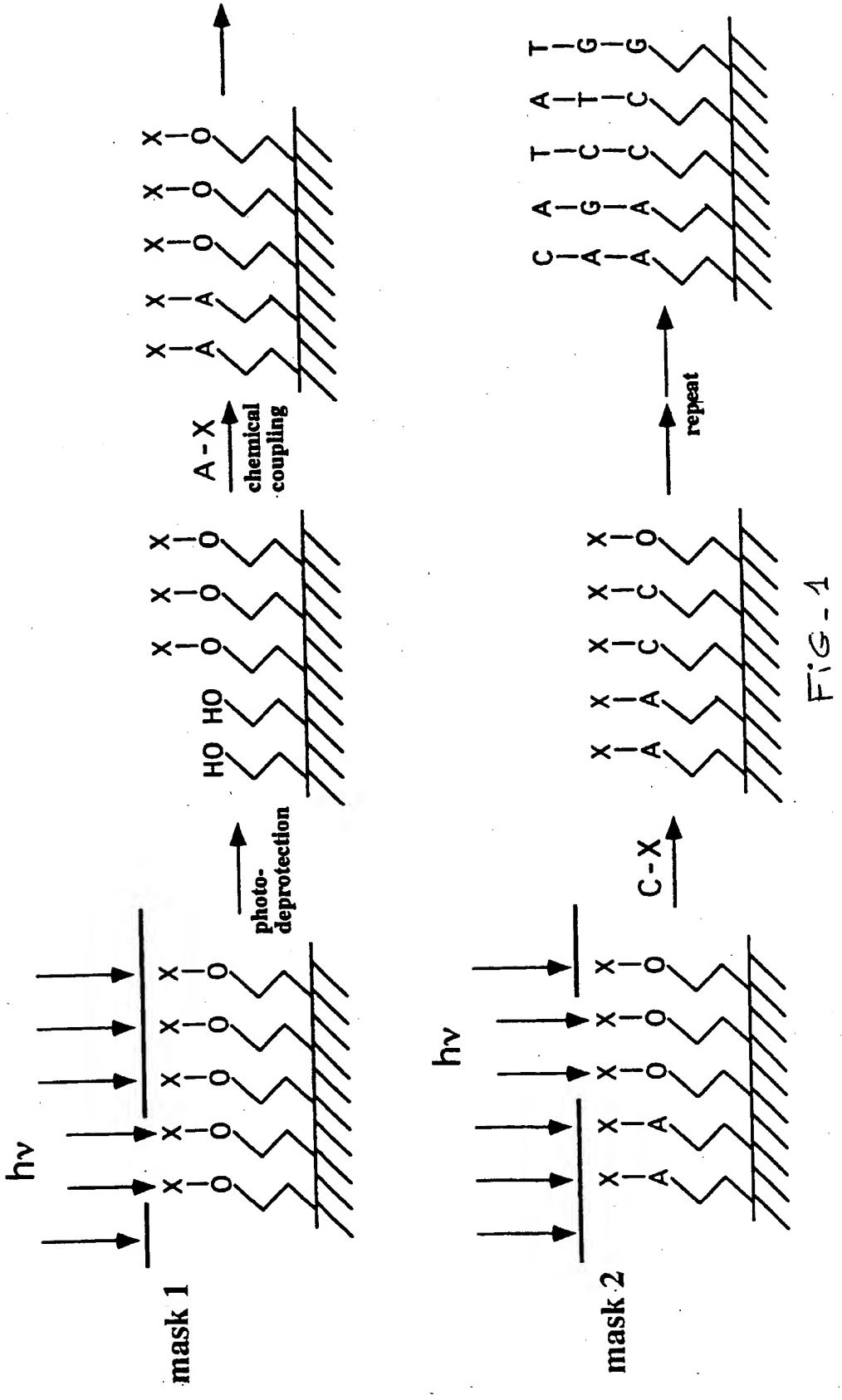
Light-Directed Synthesis of DNA Chips
hν
mask 1
photo-deprotection
A - X
chemical coupling
repeat
C-X
hν
mask 2

FIG - 1

Rsal Digestion of a Variable Density ssDNA Chip
Primer-Extended with Fluorescein-12-dATP

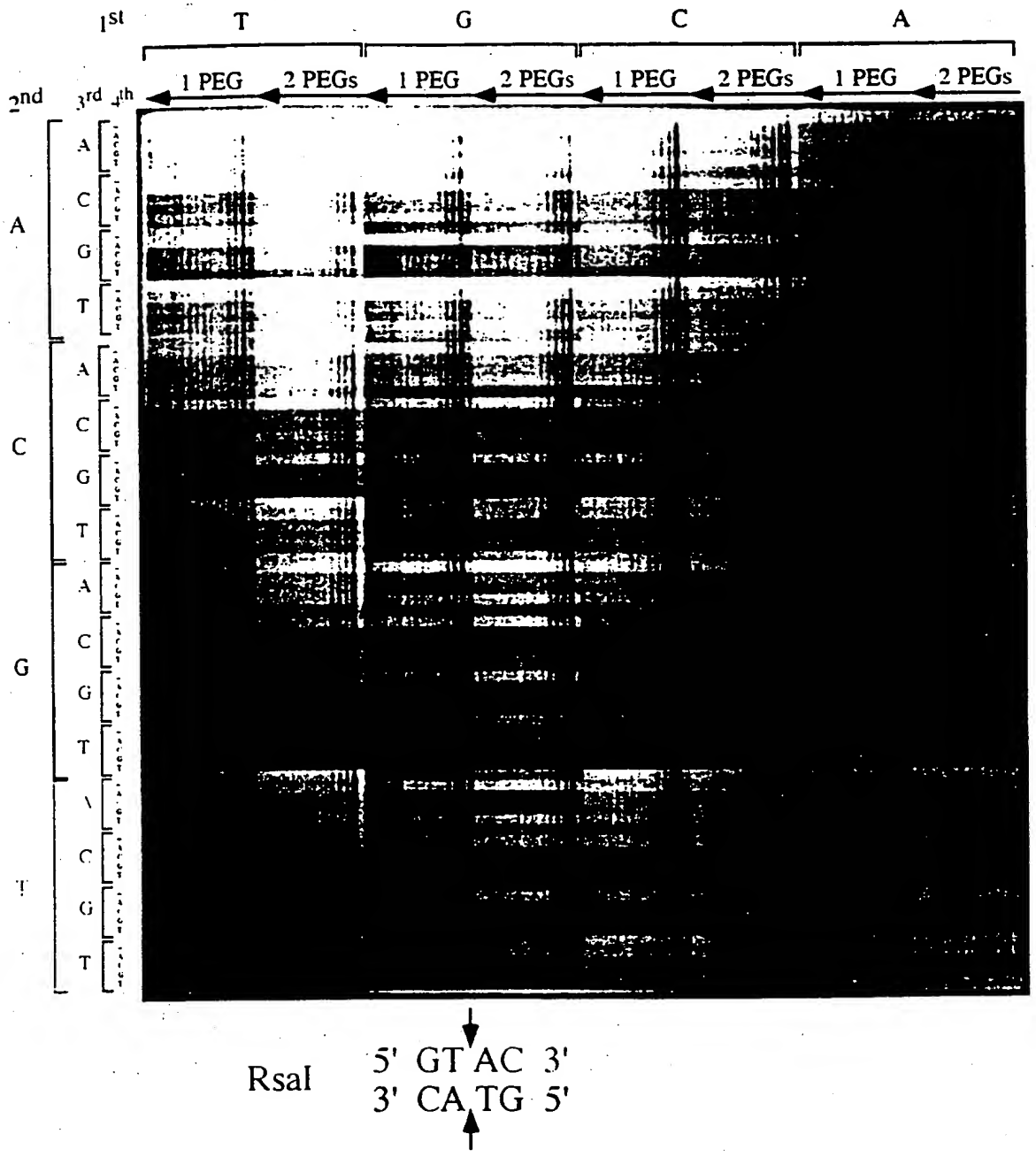

FIG. 2

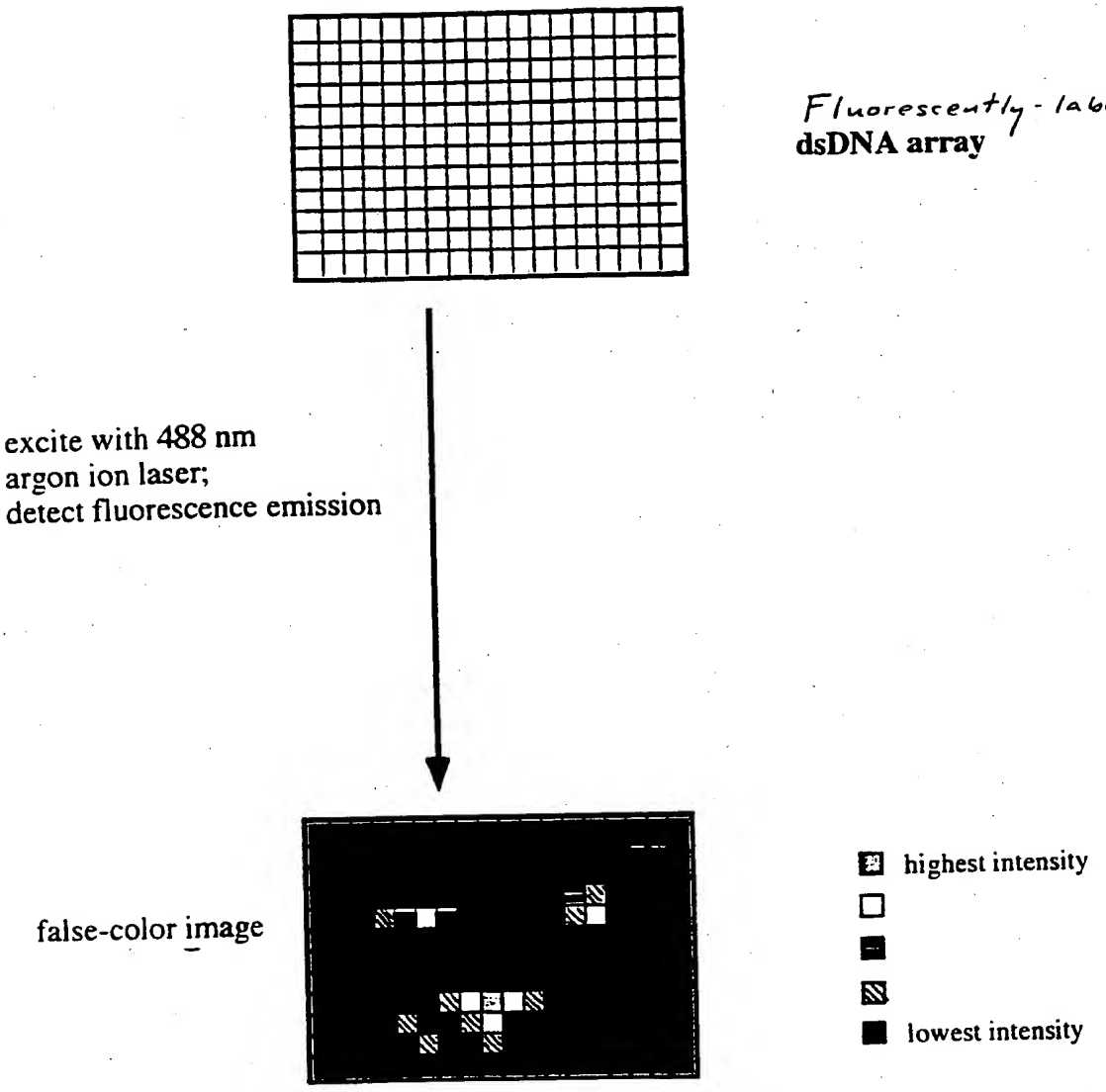
Fluorescently-labeled
dsDNA array
excite with 488 nm
argon ion laser;
detect fluorescence emission
false-color image
highest intensity
lowest intensity

FIG. 3

RsaI Digestion of a Fixed Density Double-Stranded DNA Chip with a Variable Spacer Length of 0 to 14 bp Between the Half-Sites

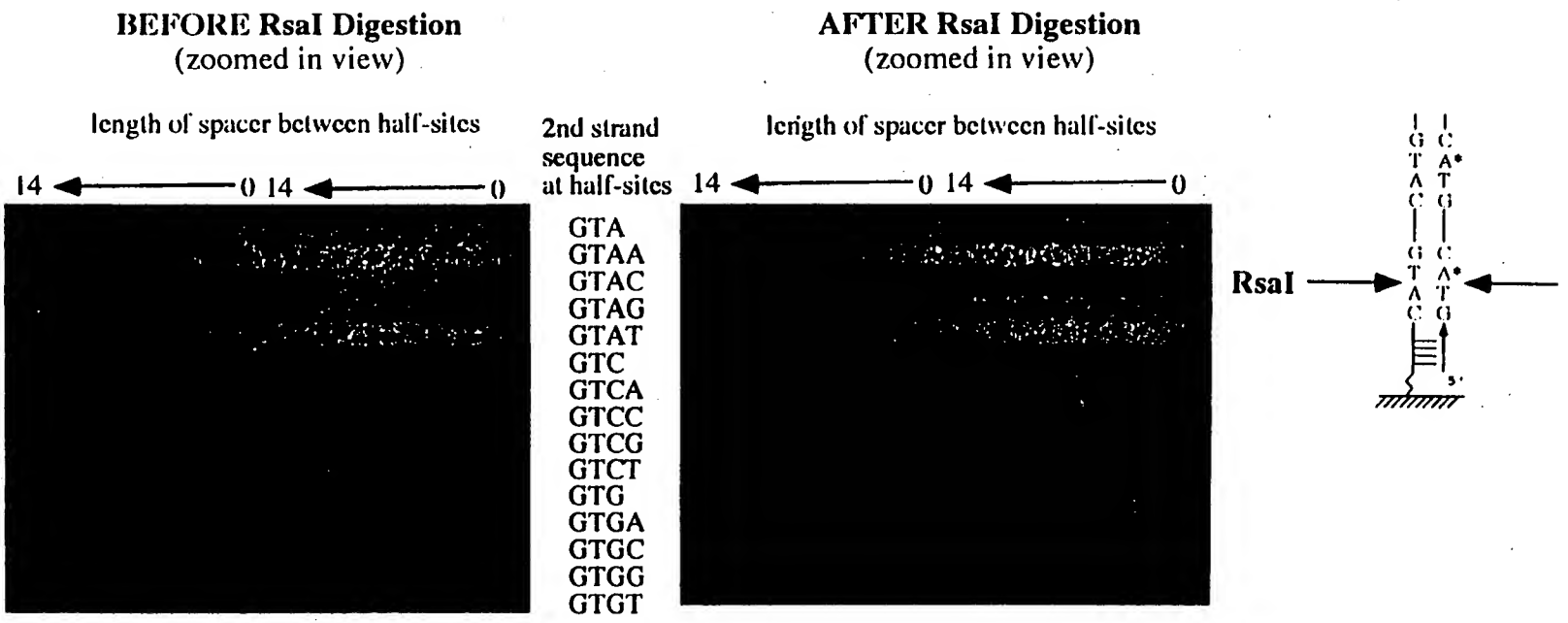

FIG. 4

DNA Chip Binding Site Assay

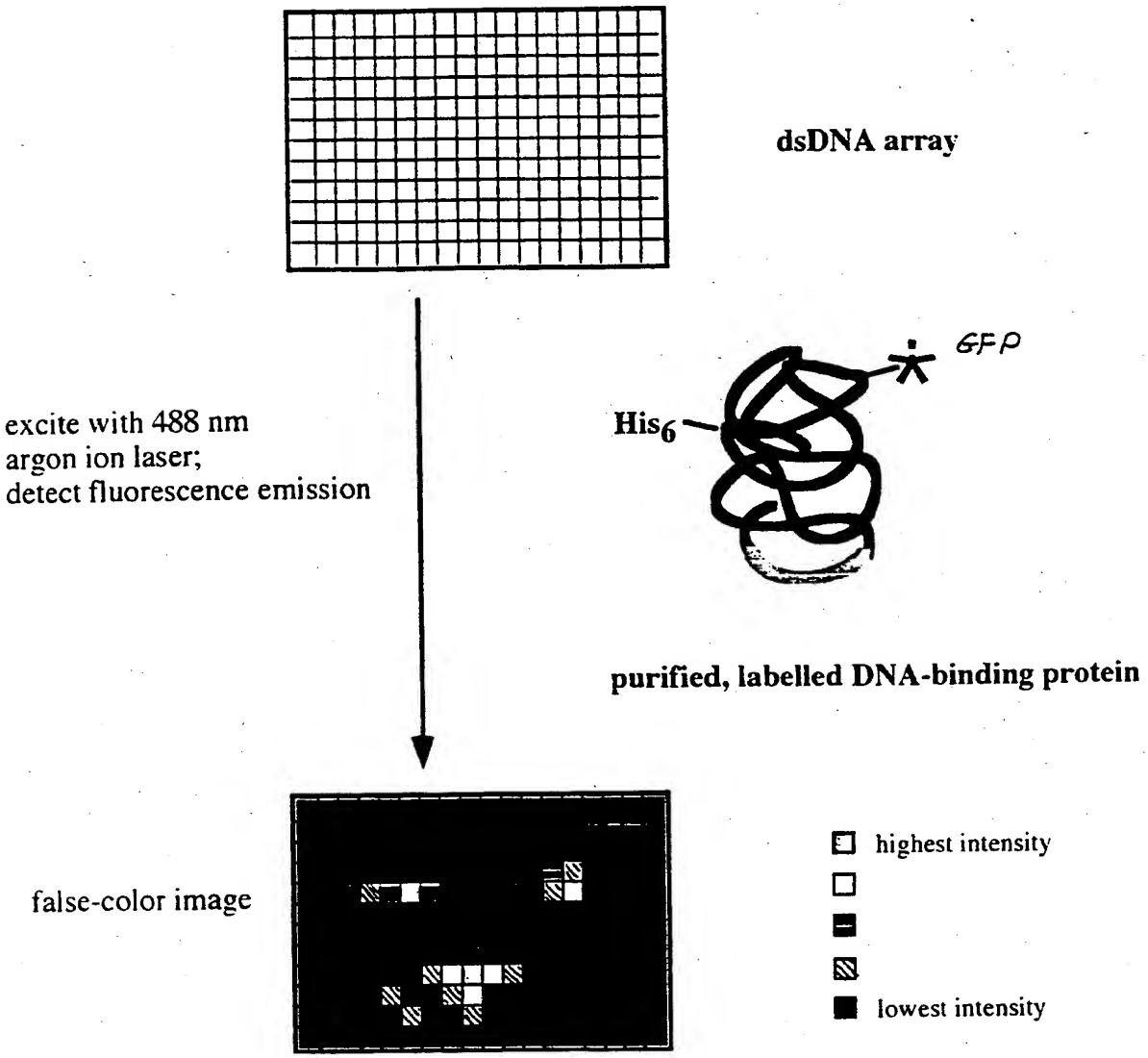

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US98/15408

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; C07H 21/02
US CL :435/6; 536/23.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

cas, biosis

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5,556,752 A (LOCKHART et al) 17 September 1996, see entire document.	1-27
Y	SCOTT et al. Searching for peptide ligands with an Epitope Library. Science, 27 July 1990. Vol. 249, pages 386-390, see entire article.	1-27
Y	LAM et al. A new type of synthetic peptide library for identifying ligand-binding activity. Nature, 07 November 1991, Vol 354, pages 82-84, see entire article.	1-27
Y	BRENNER et al. Encoded combinatorial chemistry. PNAS, June 1992. Vol 86, pages 5381-5383, see entire article.	1-27

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search
31 AUGUST 1998

Date of mailing of the international search report
28 OCT 1998

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703) 305-3230

Authorized officer
EGGERTON CAMPBELL
Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*